

# The Philosophy of Philosophy

## The Blackwell/Brown Lectures in Philosophy

Series Editor: Ernest Sosa, Brown University

The Blackwell/Brown Lectures in Philosophy present compact books distilling cutting-edge research from across the discipline. Based on public lectures presented at Brown University, the books in the series are by established scholars of the highest caliber, presenting their work in a clear and concise format.

1. *Semantic Relationism* by Kit Fine
2. *The Philosophy of Philosophy* by Timothy Williamson

# The Philosophy of Philosophy

Timothy Williamson



**Blackwell**  
Publishing

© 2007 by Timothy Williamson

BLACKWELL PUBLISHING  
350 Main Street, Malden, MA 02148-5020, USA  
9600 Garsington Road, Oxford OX4 2DQ, UK  
550 Swanston Street, Carlton, Victoria 3053, Australia

The right of Timothy Williamson to be identified as the Author of this Work has been asserted in accordance with the UK Copyright, Designs, and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs, and Patents Act 1988, without the prior permission of the publisher.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks, or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs, and Patents Act 1988, without the prior permission of the publisher.

First published 2007 by Blackwell Publishing Ltd

1 2007

*Library of Congress Cataloguing-in-Publication Data*

Williamson, Timothy.

    The philosophy of philosophy / Timothy Williamson.  
    p. cm. — (The Blackwell/Brown lectures in philosophy ; 2)

    Includes bibliographical references and index.

    ISBN 978-1-4051-3396-8 (pbk. : alk. paper) — ISBN 978-1-4051-3397-5  
(hardcover : alk. paper) 1. Philosophy. I. Title.

B53.W495 2007  
101—dc22

2007019838

A catalogue record for this title is available from the British Library.

Set in 10.5 on 13 pt Sabon  
by SNP Best-set Typesetter Ltd., Hong Kong  
Printed and bound in the United Kingdom  
by TJ International Ltd, Padstow, Cornwall

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy, and which has been manufactured from pulp processed using acid-free and elementary chlorine-free practices. Furthermore, the publisher ensures that the text paper and cover board used have met acceptable environmental accreditation standards.

For further information on  
Blackwell Publishing, please visit our website:  
[www.blackwellpublishing.com](http://www.blackwellpublishing.com)

To my children Alice, Conrad, and Arno

# Contents

---

<i>Preface</i>	<i>ix</i>
<i>Acknowledgments</i>	<i>xi</i>
Introduction	1
1 The Linguistic Turn and the Conceptual Turn	10
2 Taking Philosophical Questions at Face Value	23
3 Metaphysical Conceptions of Analyticity	48
4 Epistemological Conceptions of Analyticity	73
5 Knowledge of Metaphysical Modality	134
6 Thought Experiments	179
7 Evidence in Philosophy	208
8 Knowledge Maximization	247
Afterword Must Do Better	278
Appendix 1 Modal Logic within Counterfactual Logic	293
Appendix 2 Counterfactual Donkeys	305
<i>Bibliography</i>	309
<i>Index</i>	322

# Preface

---

This book grew out of a sense that contemporary philosophy lacks a self-image that does it justice. Of the self-images that philosophy inherited from the twentieth century, the most prominent – naturalism, the linguistic turn, postmodern irony, and so on – seemed obviously inadequate to most of the most interesting work in contemporary philosophy: as descriptions, false when bold, uninformative when cautious. Less prominent alternatives too seemed implausible or ill-developed. Although an adequate self-image is not a precondition of all virtue, it helps. If philosophy misconceives what it is doing, it is likely to do it worse. In any case, an adequate self-image is worth having for its own sake; we are not supposed to be leading the unexamined life. This is my attempt to do better.

I considered using the phrase “philosophical method” in the title, but decided against on the grounds that it seemed to promise something more like a recipe for doing philosophy than I believe possible. When asked for advice on some occasion, the Duke of Wellington is said to have replied “Sir, you are in a devilish awkward predicament, and must get out of it as best you can.” My advice would be scarcely more useful. At the crucial point, I can only say “Use your judgment.” The primary task of the philosophy of science is to understand science, not to give scientists advice. Likewise, the primary task of the philosophy of philosophy is to understand philosophy, not to give philosophers advice – although I have not rigorously abstained from the latter.

I also rejected the word “metaphilosophy.” The philosophy of philosophy is automatically part of philosophy, just as the philosophy of anything else is, whereas metaphilosophy sounds as though it might try to look down on philosophy from above, or beyond. One

reason for the survival of implausible self-images of philosophy is that they have been insufficiently scrutinized as pieces of philosophy. Passed down as though they were platitudes, they often embody epistemologically or logically naïve presuppositions. The philosophy of philosophy is no easier than the philosophy of science. And like the philosophy of science, it can only be done well by those with some respect for what they are studying.

The book makes no claim to comprehensiveness. For example, it does not engage in detail with critics of analytic philosophy who do not engage with it in detail. I preferred to follow a few lines of thought that I found more rewarding. I hope that philosophy as I have presented it seems worth doing and not impossibly difficult. At any rate, I enjoy it.

# Acknowledgments

---

My three Blackwell/Brown lectures, given at Brown University in September 2005, constituted the occasion for the book, although the material has evolved considerably since then. I thank both Blackwell Publishing and Brown University for the invitation and their generous hospitality. Jeff Dean at Blackwell has been a helpful and supportive editor.

My further debts of gratitude are huge. An earlier version of some of the material was presented as the Jack Smart Lecture at the Australian National University in July 2005. Various later versions were presented as four Anders Wedberg Lectures at the University of Stockholm in April 2006, where the commentators were Kathrin Glüer-Pagin, Sören Häggqvist, Anna-Sara Malmgren and Åsa Wikforss, as eight José Gaos Lectures at the Instituto de Investigaciones Filosóficas of the Universidad Nacional Autónoma de México in September–October 2006, and as three Carl G. Hempel Lectures at Princeton University in December 2006. Other occasions on which the material in one form or another came under scrutiny included a week-long graduate course at the University of Bologna in May–June 2005, a week-long Kompaktseminar at the University of Heidelberg in February 2006, three lectures I gave as the Townsend Visitor in Philosophy at the University of California, Berkeley, in September 2006, a lecture and workshop at the University of Munich in June 2005, two lectures I gave as Tang Chun-I Visiting Professor at the Chinese University of Hong Kong in March 2007, and lectures at a graduate conference on epistemology at the University of Rochester in September 2004, where Richard Feldman was the commentator, the University of Arizona, Tucson, and the University of California, Los Angeles, and a meeting of the Aristotelian Society (my Presidential

Address) in October 2004, a workshop on the epistemology of philosophy at the University of Bristol in May 2005, a conference on philosophical methodology at the Research School of Social Sciences at the Australian National University in July 2005, a conference on philosophical knowledge in Erfurt, and at Rutgers University in September 2005, the University of Warwick in November 2005, an Arché workshop on modality at the University of St Andrews in December 2005, a workshop on metaphysics at the University of Nottingham in January 2006, the first conference of the Dutch-Flemish Society for Analytic Philosophy at the University of Amsterdam and the University of Leeds in March 2006, the Universities of Turin and Milan, the “Is there anything wrong with Wittgenstein?” conference in Reggio Emilia and the third conference of the Portuguese Society for Analytic Philosophy at the University of Lisbon in June 2006, the Joint Session of the Aristotelian Society and the Mind Association at the University of Southampton in July 2006 (my address as President of the Mind Association), the GAP.6 conference of the German Society for Analytic Philosophy and the subsequent workshop on Implicit Definitions and A Priori Knowledge, where Frank Hofmann was the commentator, at the Free University of Berlin in September 2006, the University of Santiago de Compostela in November 2006, the Massachusetts Institute of Technology and the Eastern Division meeting of the American Philosophical Association, for which Gillian Russell was the commentator, in December 2006, the Royal Institute of Philosophy and the University of Calgary in February 2007, and the University of Cambridge in June 2007. I presented still earlier versions of the ideas at a workshop on intuition and epistemology at the University of Fribourg, where Manuel García Carpintero was the commentator, a conference on modalism and mentalism in contemporary epistemology hosted by Aarhus University at the Carlsberg Academy in Copenhagen, a conference in the Centre for Advanced Studies at the Norwegian Academy of Science and Letters in Oslo, which also hosted me for a term of leave in the summer of 2004, a workshop at the University of Amiens on John Cook Wilson and Oxford realism, a conference on externalism, phenomenology, and understanding in memory of Greg McCulloch at the Institute of Philosophy in the University of London’s School of Advanced Study, a summer school on epistemology at the Sorbonne, and a conference on meaning and truth at St Andrews, and talks at the universities

of Bilkent, Edinburgh, Michigan, Minnesota, Padua, Rijeka, and Stirling. Most of the material has also been presented in classes and discussion groups at Oxford. Much of the development of themes in this book was provoked by reflection on the questions and objections raised on these occasions. It would be hopeless to try to enumerate the questioners and objectors, but they may be able to trace their influence.

Those who have helped with discussion or written comments outside the occasions above include Alexander Bird, Stephan Blatti, Davor Bodrožić, Berit Brogaard, Earl Conee, Keith DeRose, Dorothy Edgington, Pascal Engel, Tamar Szabó Gendler, Olav Gjelsvik, John Hawthorne, Thomas Kroedel, Brian Leftow, Brian Leiter, Peter Lipton, Ofra Magidor, Mike Martin, Nenad Miščević, Michael Pendlebury, Oliver Pooley, Gonzalo Rodriguez-Pereyra, Helge Rückert, Joe Salerno, Laura Schroeter, Nico Silins, Jason Stanley, Scott Sturgeon, Hamid Vahid, Alberto Voltolini, and Ralph Wedgwood. John Hawthorne, Joshua Schechter, and two referees read the book in manuscript and provided comments on which I drew extensively during the final revisions.

That list of acknowledgements is undoubtedly incomplete: special thanks to those who have been undeservedly omitted.

The book is based on a series of articles in which earlier versions of the ideas were formulated, although hardly any pages have survived completely unchanged. Chapters 1 and 2 derive from “Past the Linguistic Turn?,” in *The Future for Philosophy*, edited by Brian Leiter (Oxford: Oxford University Press, 2004), pp. 106–28. Most of Chapter 3 is new. The first section of Chapter 3 and much of Chapter 4 constitute an expanded version of “Conceptual Truth,” *Aristotelian Society*, supplementary volume 80 (2006), pp. 1–41, with much new material (for example, on tacit knowledge and on normative conceptions of analyticity); the germ is to be found in “Understanding and Inference,” *Aristotelian Society*, supplementary volume 77 (2003), pp. 249–93. Chapters 5 and 6 derive from an initial sketch in my Presidential Address to the Aristotelian Society, “Armchair Philosophy, Metaphysical Modality and Counterfactual Thinking,” *Proceedings of the Aristotelian Society*, volume 105 (2005): 1–23. An intermediate step on the way to Chapter 5 was “Philosophical Knowledge and Knowledge of Counterfactuals,” *Grazer Philosophische Studien*, volume 74 (2007): 89–123, also

appearing as *Philosophical Knowledge – Its Possibility and Scope*, edited by Christian Beyer and Alex Burri (Amsterdam: Rodopi, 2007), the proceedings of the Erfurt conference on philosophical knowledge. Chapters 7 and 8 derive from “Philosophical ‘Intuitions’ and Scepticism about Judgement,” *Dialectica* 58 (2004), pp. 109–53; the volume constitutes the proceedings of the workshop on intuition and epistemology at the University of Fribourg, Switzerland, in November 2002 (the talk I gave there is not recognizable in this book; I gave it to make myself think seriously about the topic). Chapter 7 in particular has been greatly expanded; sections 1 and 7 are new; the probabilistic material in section 4 is expanded from pp. 683–5 of “Knowledge and Scepticism,” *The Oxford Handbook of Contemporary Philosophy*, edited by Frank Jackson and Michael Smith, (Oxford: Oxford University Press, 2005), pp. 681–700. The Afterword is a slightly modified version of “Must Do Better,” in *Truth and Realism*, edited by Patrick Greenough and Michael Lynch (Oxford: Oxford University Press: 2006), pp. 177–87; the volume constitutes the proceedings of the St Andrews conference on meaning and truth.

Thanks above all to my wife Ana, who does not let me forget what matters.

# Introduction

---

What can be pursued in an armchair?

Every armchair pursuit raises the question whether its methods are adequate to its aims. The traditional methods of philosophy are armchair ones: they consist of thinking, without any special interaction with the world beyond the chair, such as measurement, observation or experiment would typically involve. To do justice to the social and not solely individual nature of philosophy, as a dialectic between several parties, we should add speaking and listening to thinking, and allow several armchairs, within earshot of each other, but methodologically that brings philosophy little closer to the natural sciences. For good or ill, few philosophers show much appetite for the risky business of making predictions and testing them against observation, whether or not their theories in fact have consequences that could be so tested. Without attempting to define the terms precisely, we may put the difference to a first approximation thus: the current methodology of the natural sciences is *a posteriori*; the current methodology of philosophy is *a priori*. What should we make of this difference?

Opposite reactions are possible. *Crude rationalists* regard philosophy's *a priori* methodology as a virtue. According to them, it makes philosophical results especially reliable, because immune from perceptual error. *Crude empiricists* regard philosophy's *a priori* methodology as a vice. According to them, it makes philosophical results especially unreliable, because immune from perceptual correction.

Few contemporary philosophers have the nerve to be crude rationalists. Given the apparent absence of a substantial body of agreed results in philosophy, crude rationalism is not easy to maintain. Many contemporary philosophers have some sympathy for crude empiri-

cism, particularly when it goes under the more acceptable name of “naturalism.” However, that sympathy sometimes has little effect on their philosophical practice: they still philosophize in the grand old manner, merely adding naturalism to their list of *a priori* commitments.

A subtler response to naturalism, or empiricism, is to scale down the ambitions of philosophy. Holding fixed its *a priori* methodology, one asks what it could be good for. Not for answering ordinary factual questions, it is claimed: that is best left to the natural sciences with their *a posteriori* methodology. Nevertheless, what we already have in the armchair is the intellectual equipment we bring to *a posteriori* inquiry, our conceptual or linguistic competence. Perhaps philosophy can find some sort of legitimate employment by investigating, from within, what we bring to inquiry. Rather than trying to answer ordinary factual questions, it seeks to understand the very possibility of asking them – in some way, yet to be properly specified, that does not involve asking ordinary factual questions about the possibility of asking ordinary factual questions. The “linguistic turn” in twentieth-century philosophy comprises a variety of attempts in that general spirit. Since confinement to an armchair does not deprive one of one’s linguistic competence, whatever can be achieved through exercise of that competence and reflection thereon will be a feasible goal for philosophy. If one regards thought as constituting a more fundamental level of analysis than language, one may generalize the linguistic turn to the “conceptual turn,” and consider what can be achieved through exercise of our conceptual competence and reflection thereon, but the outcome will be broadly similar: philosophical questions turn out to be in some sense conceptual questions.

Crude rationalists, crude empiricists, and linguistic or conceptual philosophers (those who take the linguistic or conceptual turn) share a common assumption: that the *a priori* methodology of philosophy is profoundly unlike the *a posteriori* methodology of the natural sciences; it is no mere difference between distinct applications of the same underlying methodology. One apparently distinctive feature of current methodology in the broad tradition known as “analytic philosophy” is the appeal to *intuition*. Crude rationalists postulate a special knowledge-generating faculty of rational intuition. Crude empiricists regard “intuition” as an obscurantist term for folk prejudice, a psychological or social phenomenon that cannot legitimately

constrain truth-directed inquiry. Linguistic or conceptual philosophers treat intuitions more sympathetically, as the deliverances of linguistic or conceptual competence. Of course, the appeal to intuitions also plays a crucial role in the overt methodology of other disciplines too, such as linguistics.

One main theme of this book is that the common assumption of philosophical exceptionalism is false. Even the distinction between the *a priori* and the *a posteriori* turns out to obscure underlying similarities. Although there are real methodological differences between philosophy and the other sciences, as actually practiced, they are less deep than is often supposed. In particular, so-called intuitions are simply judgments (or dispositions to judgment); neither their content nor the cognitive basis on which they are made need be distinctively philosophical. In general, the methodology of much past and present philosophy consists in just the unusually systematic and unrelenting application of ways of thinking required over a vast range of non-philosophical inquiry. The philosophical applications inherit a moderate degree of reliability from the more general cognitive patterns they instantiate. Although we cannot prove, from a starting-point a sufficiently radical skeptic would accept, that those ways of thinking are truth-conducive, the same holds of *all* ways of thinking, including the methods of natural science. That is the skeptic's problem, not ours. By more discriminating standards, the methodology of philosophy is not in principle problematic.

Some may wonder whether philosophy *has* a method to be studied, especially if it is as methodologically undistinctive as just suggested. Forget the idea of a single method, employed in all and only philosophical thinking. Still, philosophers use methods of various kinds: they philosophize in various ways. A philosophical community's methodology is its repertoire of such methods. The word "method" here carries no implication of a mechanically applicable algorithm, guaranteed to yield a result within a finite time. On this loose understanding of what a methodology is, it is disingenuous for a philosopher to claim to have none.

Another main theme of this book is that the differences in subject matter between philosophy and the other sciences are also less deep than is often supposed. In particular, few philosophical questions are conceptual questions in any distinctive sense, except when philosophers choose to ask questions about concepts, as they may but need

not do. Philosophical questions are those philosophers are disposed to ask, which in turn tend, unsurprisingly, to be those more amenable to philosophical than to other ways of thinking; since the philosophical ways of thinking are not different in kind from the other ways, it is equally unsurprising that philosophical questions are not different in kind from other questions. Of course, philosophers are especially fond of abstract, general, necessary truths, but that is only an extreme case of a set of intellectual drives present to some degree in all disciplines.

In most particular cases, philosophers experience little difficulty in recognizing the difference between philosophy and non-philosophy. Being philosophers, they care about the difference, and have a professional temptation to represent it as a deep philosophical one. But just about every institutionally distinct discipline acquires a professional identity, and its practitioners experience little difficulty in recognizing the difference between what “we” do and what “they” do in most particular cases. They care about the difference, and have a professional temptation to represent it in the terms of their own discipline. But such temptations can be resisted. The distinction between the Department of Philosophy and the Department of Linguistics or the Department of Biology is clearer than the distinction between philosophy and linguistics or biology; the philosophy of language overlaps the semantics of natural languages and the philosophy of biology overlaps evolutionary theory.

The unexceptional nature of philosophy is easier to discern if we avoid the philistine emphasis on a few natural sciences, often imagined in crudely stereotyped ways that marginalize the role of armchair methods in those sciences. Not all science is natural science. Whatever crude empiricists may say, mathematics is a science if anything is; it is done in an armchair if anything is. In no useful sense are mathematical questions conceptual questions. If mathematics is an armchair science, why not philosophy too?

Most philosophers are neither crude rationalists nor crude empiricists nor, these days, linguistic or conceptual philosophers. Many would accept the theses just enunciated about the methodology and subject matter of philosophy. But a third theme of this book is that the current philosophical mainstream has failed to articulate an adequate philosophical methodology, in part because it has fallen into

the classic epistemological error of psychologizing the data. For example, our evidence is sometimes presented as consisting of our intuitions: not their content, since it is allowed that some of our intuitions may be false, but rather our psychological states of having those intuitions. We are then supposed to infer to the philosophical theory that best explains the evidence. But since it is allowed that philosophical questions are typically not psychological questions, the link between the philosophical theory of a non-psychological subject matter and the psychological evidence that it is supposed to explain becomes problematic: the description of the methodology makes the methodology hard to sustain. Again, philosophy is often presented as systematizing and stabilizing our beliefs, bringing them into reflective equilibrium: the picture is that in doing philosophy what we have to go on is what our beliefs currently are, as though our epistemic access were only to those belief states and not to the states of the world that they are about. The picture is wrong; we frequently have better epistemic access to our immediate physical environment than to our own psychology. A popular remark is that we have no choice but to start from where we are, with our current beliefs. But where we are is not only having various beliefs about the world; it is also having significant knowledge of the world. Starting from where we are involves starting from what we already know, and the goal is to know more (of course, how much more we come to know cannot be measured just by the number of propositions learnt). To characterize our method as one of achieving reflective equilibrium is to fail to engage with epistemologically crucial features of our situation. Our understanding of philosophical methodology must be rid of internalist preconceptions.

Philosophical errors distort our conception of philosophy in other ways too. Confused and obscure ideas of conceptual truth create the illusion of a special domain for philosophical investigation. Similarly, although perception clearly involves causal interaction between perceiver and perceived, crudely causal accounts of perceptual knowledge that occlude the contribution of background theory create the illusion of a contrast between world-dependent empirical beliefs and world-independent philosophical theory.

Clearly, the investigation of philosophical methodology cannot and should not be philosophically neutral. It is just more philosophy,

turned on philosophy itself. We have the philosophy of mathematics, the philosophy of physics, the philosophy of biology, the philosophy of economics, the philosophy of history; we also need the philosophy of philosophy.

The rethinking of philosophical methodology in this book involves understanding, at an appropriate level of abstraction, how philosophy is actually done. Philosophers of science know the dangers of moralizing from first principles on how a discipline should ideally be pursued without respecting how it currently is pursued; the same lesson applies to the philosophy of philosophy. The present opposition to philosophical exceptionalism is far from involving the idea that philosophers should model themselves on physicists or biologists. The denial that philosophical questions are conceptual questions is quite compatible with a heavy emphasis on issues of semantic structure in philosophical discussion, for the validity or otherwise of philosophical reasoning is often highly sensitive to delicate aspects of the semantic structure of premises and conclusion: to make our reasoning instruments more reliable, we must investigate those instruments themselves, even when they are not the ultimate objects of our concern.

That philosophy *can* be done in an armchair does not entail that it *must* be done in an armchair.<sup>1</sup> This book raises no objection to the idea that the results of scientific experiments are sometimes directly relevant to philosophical questions: for example, concerning the philosophy of time. But it is a fallacy to infer that philosophy can nowhere usefully proceed until the experiments are done. In this respect, philosophy is similar to mathematics. Scientific experiments can be relevant to mathematical questions. For instance, a physical theory may entail that there are physically instantiated counter-examples to a mathematical theory. A toy example: one can specify in physical terms what it takes to be an inscription (intended or unintended) in a given font of a proof of “ $0 = 1$ ” in a given formal system of Peano Arithmetic; a physical theory could predict that an event of a specified physically possible type would cause there to be

<sup>1</sup> In this respect Hilary Kornblith seems to misunderstand the claim that philosophy can be done in an armchair (2006: 19). I have even dabbled in experimental philosophy myself (Bonini, Osherson, Viale and Williamson 1999).

such an inscription. Less directly, psychological experiments might in principle reveal levels of human unreliability in proof-checking that would undermine current mathematical practice. To conclude on that basis alone that mathematics should become an experimental discipline would be hopelessly naïve. In practice, most of mathematics will and should remain an armchair discipline, even though it is not in principle insulated from experimental findings, because armchair methods, specifically proof, remain by far the most reliable and efficient available. Although the matter is less clear-cut, something similar may well apply to many areas of philosophy, for instance, philosophical logic. In particular, on the account in this book, the method of conducting opinion polls among non-philosophers is not very much more likely to be the best way of answering philosophical questions than the method of conducting opinion polls among non-physicists is to be the best way of answering physical questions.

Although this book is a defense of armchair philosophy, it is not written in a purely conservative spirit. Our ideas about philosophical methodology, however inchoate, are liable to influence the methodology we actually employ; bad ideas about it are liable to tilt it in bad directions. A reasonable hypothesis is that our current methodology is good enough to generate progress in philosophy, but not by much: ten steps forward, nine steps back. Nevertheless, we can improve our performance even without radically new methods. We need to apply the methods we already have with more patience and better judgment. A small increase in accuracy of measurement may enable scientists to tackle problems previously beyond reach, because their data lacked sufficient resolution. Similarly, small improvements in accepted standards of reasoning may enable the philosophical community to reach knowledgeable agreement on the status of many more arguments. Such incremental progress in philosophical methodology is a realistic prospect, for current standards in the profession exhibit large variations significantly correlated with differences between graduate schools. Philosophical methodology can be taught – mainly by example, but fine-tuning by explicit precept and discussion also makes a difference. For instance, the level of rigor in philosophical statement and argument which Frege achieved only by genius (with a little help from his mathematical training) is now available to hundreds of graduate students every year: and we know how to do even better. That is not to imply, of course, that we must strive for maximum

rigor at all times, otherwise this impressionistic introduction would be self-defeating. At any rate, if the philosophical community has the will, it can gradually bring up a much higher proportion of practice to the standard of current best practice, and beyond. Such progress in methodology cannot be relied on to happen automatically; not all of us love the highest at first sight. Although the envisaged incremental progress lacks the drama after which some philosophers still hanker, that hankering is itself a symptom of the intellectual immaturity that helps hold philosophy back. No revelation is at hand; any improvement in accepted standards of philosophical discussion will result from collective hard work and self-discipline. One hope with which this book is written is that by contributing to the current tendency towards increasing methodological self-consciousness in philosophy it will play some role, however indirect, in raising those standards. Philosophizing is not like riding a bicycle, best done without thinking about it – or rather: the best cyclists surely *do* think about what they are doing.

This book is an essay. It makes no claim to comprehensiveness. It does not attempt to compile a list of philosophical methods, or of theories about philosophical methods. It touches on historical matters only glancingly. Instead, it explores some interrelated issues that strike me as interesting and not well understood. It starts by inquiring into the nature of philosophical questions. It proceeds in part by detailed case studies of particular examples. Since all examples have their own special characteristics, generalizations from them must be tentative. But many long-standing misconceptions in philosophy are helped to survive by an unwillingness to look carefully and undogmatically at examples, sometimes protected by a self-righteous image of oneself and one's friends as the only people who do look carefully and undogmatically at examples (some disciples of the later Wittgenstein come to mind).

It is difficult to displace one philosophical picture except by another. Although discussion of philosophical methodology is itself part of philosophy, it is less often conducted with a clear view of the theoretical alternatives than is usual in philosophy. David Lewis once wrote that “what we accomplish in philosophical argument” is to “measure the price” of maintaining a philosophical claim; when his remark is cited as an obvious truth, it tends not to be noticed that it too is subject to philosophical argument, and has its price – not least

the danger of infinite regress, since claims about the price of maintaining a philosophical claim are themselves subject to philosophical argument.<sup>2</sup> Another hope for this book is that it will clarify an alternative to widespread assumptions about the nature of philosophy.

<sup>2</sup> See his 1983a: x. Lewis himself gives a brief philosophical argument for his claim about measuring the price, based on the premise that “[o]ur ‘intuitions’ are simply opinions,” against a foundationalist alternative. He also qualifies the claim, allowing that Gödel and Gettier may have conclusively refuted philosophical theories, and that perhaps the price of a philosophical claim “is something we can settle more or less conclusively.”

# The Linguistic Turn and the Conceptual Turn

---

*The Linguistic Turn* is the title of an influential anthology edited by Richard Rorty, published in 1967. He credited the phrase to Gustav Bergmann (Bergmann 1964: 3; Rorty 1967: 9). In his introduction, Rorty (1967: 3) explained:

The purpose of the present volume is to provide materials for reflection on the most recent philosophical revolution, that of linguistic philosophy. I shall mean by “linguistic philosophy” the view that philosophical problems are problems which may be solved (or dissolved) either by reforming language, or by understanding more about the language we presently use.

“The linguistic turn” has subsequently become the standard vague phrase for a diffuse event – some regard it as *the* event – in twentieth-century philosophy, one not confined to signed-up linguistic philosophers in Rorty’s sense. For those who took the turn, language was somehow the central theme of philosophy.

The word “theme” is used with deliberate vagueness. It does not mean “subject matter,” for the linguistic turn was not the attempted reduction of philosophy to linguistics. The theme of a piece of music is not its subject matter. Those who viewed philosophy as an activity of dispelling confusions of linguistic origin did not see it as having a subject matter in the sense in which a science has a subject matter. But merely to regard linguistic analysis as one philosophical method among many is not yet to have taken the linguistic turn, for it is not yet to regard language as central. We will be more precise below.

There is an increasingly widespread sense that the linguistic turn is past. We will ask how far the turn has been, or should be, reversed.

Language has been regarded as central to philosophy in many different ways, which cannot all be treated together. A history of the many different forms that the linguistic turn took would be a history of much of twentieth-century philosophy. That is a task for another book, by another author. Self-indulgently, I will use a thin slice through history to introduce the contemporary issues by briefly considering some of my predecessors in the Wykeham Chair of Logic at Oxford.

A. J. Ayer was the first holder of the Chair to take the linguistic turn.<sup>1</sup> In 1936, back from Vienna and its Circle but not yet in the Chair, he announced an uncompromisingly formal version of linguistic philosophy:

[T]he philosopher, as an analyst, is not directly concerned with the physical properties of things. He is concerned only with the way in which we speak about them. In other words, the propositions of philosophy are not factual, but linguistic in character – that is, they do not describe the behaviour of physical, or even mental, objects; they express definitions, or the formal consequences of definitions. (Ayer 1936: 61–2)

Ayer traced his views back ultimately to the empiricism of Berkeley and Hume (Ayer 1936: 11). His contrast between definitions of words and descriptions of objects is, roughly, the linguistic analogue of Hume's contrast between relations of ideas and matters of fact. For an empiricist, the *a priori* methods of philosophy cannot provide us with knowledge of synthetic truths about matters of fact ("the behaviour of physical, or even mental, objects"); they yield only analytic truths concerning relations of ideas ("definitions, or the formal consequences of definitions"). A rather traditional empiricism later overshadowed the linguistic theme in Ayer's work.

Ayer was the predecessor of Sir Michael Dummett in the Wykeham Chair. Dummett gave a much-cited articulation of the linguistic turn, attributing it to Frege:

Only with Frege was the proper object of philosophy finally established: namely, first, that the goal of philosophy is the analysis of the

<sup>1</sup> Ayer's three immediate predecessors were John Cook Wilson, H. H. Joachim and H. H. Price.

structure of *thought*; secondly, that the study of *thought* is to be sharply distinguished from the study of the psychological process of *thinking*; and, finally, that the only proper method for analysing thought consists in the analysis of *language*. . . . [T]he acceptance of these three tenets is common to the entire analytical school. (Dummett 1978: 458)

On this view, thought is essentially expressible (whether or not actually expressed) in a public language, which filters out the subjective noise, the merely psychological aspects of thinking, from the inter-subjective message, that which one thinks. Dummett's own corpus constitutes one of the most imposing monuments of analytic philosophy as so defined. Unlike Ayer, he does not describe philosophical claims as definitions. Unlike Rorty, he characterizes the linguistic turn as involving distinctive claims about the subject matter of philosophy, not only about its method. On Dummett's view, Frege's insight replaced epistemology by philosophy of language as first philosophy. But this methodological innovation is supposed to be grounded in the account of the proper object of philosophy.

Elsewhere, Dummett makes clear that he takes this concern with language to be what distinguishes “analytical philosophy” from other schools (1993: 4). His account of its inception varies slightly. At one points (1993: 5), he says: “[A]nalytical philosophy was born when the ‘linguistic turn’ was taken. This was not, of course, taken uniformly by any group of philosophers at any one time: but the first clear example known to me occurs in Frege’s *Die Grundlagen der Arithmetik* of 1884.” Later (1993: 27), we read: “If we identify the linguistic turn as the starting-point of analytical philosophy proper, there can be no doubt that, to however great an extent Frege, Moore and Russell prepared the ground, the crucial step was taken by Wittgenstein in the *Tractatus Logico-Philosophicus* of 1922.” Presumably, in Frege the linguistic turn was a fitful insight, in Wittgenstein, a systematic conception.

That “analytical philosophers” in Dummett’s sense coincide with those usually classified as such is not obvious. Some kind of linguistic turn occurred in much of what is usually called “continental [supposedly non-analytic] philosophy.” That Jacques Derrida did not subscribe in his own way to Dummett’s three tenets is unclear: if some stretching of terms is required, it is for the later Wittgenstein

too. Conversely, Bertrand Russell did not subscribe to the three tenets, although often cited as a paradigm “analytical philosopher.” Over the past 20 years, fewer and fewer of those who would accept the label “analytic philosophy” for their work would also claim to take the linguistic turn (I am not one of those few). Even philosophers strongly influenced by Dummett, such as Gareth Evans, Christopher Peacocke, and John Campbell, no longer give language the central role he describes. For Dummett, they belong to a tradition that has grown out of “analytical philosophy” without themselves being “analytical philosophers” (1993: 4–5). In effect, they aimed to analyze thought directly, without taking a diversion through the analysis of language. In the 1980s it became commonplace in some circles to suggest that the philosophy of mind had displaced the philosophy of language in the driving seat of philosophy.

For philosophers of mind who accepted Jerry Fodor’s (1975) influential hypothesis of a language of thought, the priority of thought to public language did not imply the priority of thought to all language, since thought itself was in a language, the brain’s computational code. In principle, someone might combine that view with Dummett’s three tenets of analytic philosophy, contrary to Dummett’s intention; he did not mean a private language. Moreover, the first-personal inaccessibility of the language of thought makes such a version of the linguistic turn methodologically very different from the traditional ones.

For those who deny the methodological priority of language to thought, the minimal fallback from Dummett’s three tenets is to reject the third but maintain the first two. They assert that the goal of philosophy is the analysis of the structure of thought, and that the study of thought is to be sharply distinguished from the study of the psychological process of thinking, but deny that the only proper method for analysing thought consists in the analysis of language. If thought has constituents, we may call them “concepts.” On this view, concepts take the place of words in Dummett’s analytical philosophy.

In practice, linguistic philosophers were often happy enough to speak of concepts rather than words, for they regarded a concept as what synonymous expressions had in common; their primary interest was in the features common to synonyms, not in the differences between them. It is therefore not too misleading to describe as *conceptual philosophers* those who accept Dummett’s first two tenets –

that the goal of philosophy is the analysis of the structure of thought, and that the study of thought is to be sharply distinguished from the study of the psychological process of thinking – whether or not they accept the third. We may also describe them as doing *conceptual philosophy*, and as having taken the *conceptual turn*.

The conceptual turn constitutes a much broader movement than the linguistic turn. It is neutral over the relative priority of language and thought. We think and talk about things – truly or falsely depending on whether they are or are not as we think or say they are. The aboutness of thought and talk is their *intentionality*; the conceptual turn puts intentionality at the centre of philosophy. This terminology indicates how little the conceptual turn is confined to what would ordinarily be called “analytic philosophy.” The phenomenological tradition may constitute another form of the conceptual turn. In the hermeneutic study of interpretation and various shades of postmodernist discourse about discourse the conceptual turn takes a more specifically linguistic form.

Have we stretched our terms so far that all philosophy is conceptual philosophy? No. On a natural view, concepts constitute only a small fraction of a largely mind-independent reality. That the goal of philosophy is in some sense to analyze that small fraction is no platitude. To put it very schematically, let *absolute idealism about the subject matter of philosophy* be the view that philosophy studies only concepts, in contrast to *ontological absolute idealism*, the wilder view that only concepts exist.<sup>2</sup> Although absolute idealism about the subject matter of philosophy does not entail ontological absolute idealism, why should we accept absolute idealism about the subject matter of philosophy if we reject ontological absolute idealism? Of course, we might reject absolute idealism about the subject matter of philosophy while nevertheless holding that the correct method for philosophy is to study its not purely conceptual subject matter by studying concepts of that subject matter. This methodological claim will be considered later; for present purposes, we merely note how much weaker it is than those formulated by Ayer and Dummett.

The claim that concepts constitute only a small fraction of reality might be opposed on various grounds. Recall that concepts were

<sup>2</sup> The “absolute” is to distinguish these forms of idealism from the corresponding “subjective” forms, in which concepts are replaced by psychological processes.

defined as the constituents of thought. If thought consists of Russellian propositions, complexes of the objects, properties, relations, and other elements of reality the proposition is about, then those objects, properties, relations, and other elements of reality are by definition concepts. In that case, ontological absolute idealism may be a triviality, because whatever exists is a constituent of various Russellian propositions, and thereby counts as a concept. However, even conceptual philosophers who accept the Russellian view of propositions will distinguish *conceptual structure*, the structure characteristic of propositions, from other sorts of structure. For example, they will analyze the atomic proposition that this crystal is translucent as the object-property complex *<this crystal, translucency>*, but they will not regard it as any of their business to analyze the structure of the crystal itself: that is chemical structure, not conceptual structure in the relevant sense, otherwise the proposition would not be atomic. Their goal for philosophy – to analyze the structure of thought – is still only to analyze one sort of structure among many. Thus one might accept the Russellian view of propositions and still oppose the conceptual turn, on the grounds that philosophy can appropriately investigate general features of nonconceptual structure too, such as the general mereological structure of physical objects.

Alternatively, take a more standard view of concepts, as something like modes of presentation, ways of thinking or speaking, or intellectual capacities. Still, the claim that concepts constitute only a small fraction of reality might be accused of violating Dummett's second tenet by confusing thought with the process of thinking. Almost everyone agrees that psychological events constitute only a small fraction of reality, but that is not yet to concede that thought in a non-psychologistic sense is similarly confined. John McDowell (1994: 27), for instance, argues:<sup>3</sup>

[T]here is no ontological gap between the sort of thing one can mean, or generally the sort of thing one can think, and the sort of thing that can be the case. When one thinks truly, what one thinks *is* what is the

<sup>3</sup> Although McDowell is sometimes classified as a “post-analytic” philosopher, he finds his own way to accept Dummett's “fundamental tenet of analytical philosophy,” that “philosophical questions about thought are to be approached through language” (1994: 125).

case. So since the world is everything that is the case . . . there is no gap between thought, as such, and the world. Of course thought can be distanced from the world by being false, but there is no distance from the world implicit in the very idea of thought.

For McDowell, the sort of thing one can think is a conceptual content: the conceptual has no outer boundary beyond which lies unconceptualized reality. He denies the accusation of idealism on the grounds that he is not committed to any contentious thesis of mind-dependence.

The sort of thing that can be the case is that a certain object has a certain property. McDowell's claim is not that the object and the property *are* concepts, but merely that we can in principle form concepts *of* them, with which to think that the object has the property. Indeed, we can in principle form many different concepts of them: we can think of the same object as Hesperus or as Phosphorus. In Fregean terms congenial to McDowell, different senses determine the same reference. He admits "an alignment of minds with the realm of sense, not with the realm of reference . . . thought and reality meet in the realm of sense" (1994: 179–80). For objects, his claim that the conceptual is unbounded amounts to the claim that any object can be thought of. Likewise for the sort of thing that can be the case: the claim is, for example, that whenever an object has a property, it can be thought, of the object and the property, that the former has the latter. But, on a coherent and natural reading of "the sort of thing that can be the case," such things are individuated coarsely, by the objects, properties, and relations that they involve. Thus, since Hesperus *is* Phosphorus, what is the case if Hesperus is bright *is* what is the case if Phosphorus is bright: the objects are the same, as are the properties. On this reading, McDowell's claim "When one thinks truly, what one thinks *is* what is the case" is false, because what one thinks is individuated at the level of sense while what is the case is individuated at the level of reference. Although McDowell's claim is true on weaker readings, they will not bear the weight his argument puts on them.

McDowell's argument in any case seems to require the premise that everything (object, property, relation, state of affairs, . . .) is thinkable. That premise is highly contentious. What reason have we to assume that reality does not contain *elusive objects*, incapable in

principle of being individually thought of? Although we can think of them collectively – for example, as elusive objects – that is not to single out any one of them in thought. Can we be sure that ordinary material objects do not consist of clouds of elusive sub-sub-atomic particles? We might know them by their collective effects while unable to think of any single one of them. The general question whether there can be elusive objects looks like a good candidate for philosophical consideration. Of course, McDowell does not intend the conceptual to be limited by the merely medical limitations of human beings, but the elusiveness may run deeper than that: the nature of the objects may preclude the kind of separable causal interaction with complex beings that isolating them in thought would require. In Fregean terminology again, a sense is a mode of presentation of a referent; a mode of presentation of something is a way of presenting it to a possible thinker, if not an actual one; for all McDowell has shown, there may be necessary limitations on thinking.<sup>4</sup> Although elusive objects belong to the same very general ontological category of objects as those we can single out, their possibility still undermines McDowell's claim that we cannot make "interesting sense" of the idea of something outside the conceptual realm (1994: 105–6). We do not know whether there actually are elusive objects. What would motivate the claim that there are none, if not some form of idealism very far from McDowell's intentions? We should adopt no conception of philosophy that on methodological grounds excludes elusive objects.<sup>5</sup>

Suppose, just for the sake of argument, that there are no elusive objects. That by itself would still not vindicate a restriction of philosophy to the conceptual, the realm of sense or thought. The practitioners of any discipline have thoughts and communicate them,

<sup>4</sup> McDowell's invocation of humility (1994: 40) addresses contingent limitations, not necessary ones.

<sup>5</sup> Mark Johnston (1993: 96–7) discusses "the Enigmas, entities essentially undetectable by us." He stipulates that they are collectively as well as individually undetectable; thus our elusive objects need not be his Enigmas. If we cannot have good evidence that there are no Enigmas, it may well be a waste of time to worry whether there are Enigmas. But it would not follow that it is a waste of time to worry whether there *can be* Enigmas. Their definition does not rule out knowledge of the possibility of such things; such knowledge may itself be philosophically useful (indeed, Johnston uses it for his philosophical purposes).

but they are rarely studying those very thoughts: rather, they are studying what their thoughts are about. Most thoughts are not about thoughts. To make philosophy the study of thought is to insist that philosophers' thoughts should be about thoughts. It is not obvious why philosophers should accept that restriction.

Even within what is usually considered analytic philosophy of mind, much work violates the two tenets of conceptual philosophy. Naturalists hold that everything is part of the natural world, and should be studied as such; many of them study thought as part of the natural world by not sharply distinguishing it from the psychological process of thinking. Those who study sensations or qualia without treating them as intentional phenomena are not usually attempting to analyze the structure of thought; their interest is primarily in the nature of the sensations or qualia themselves, not in our concepts of them. Even when the question of veridicality arises, it is not always conceded that there are structured thoughts: some philosophers claim that perception has a conceptually unstructured content that represents the environment as being a certain way. Their interest is in the nature of the nonconceptual content itself, not just in our concept of it.

Despite early hopes or fears, philosophy of mind has not come to play the organizing role in philosophy that philosophy of language once did. No single branch of philosophy does: philosophy is no more immune than other disciplines to increasing specialization. Nor is any one philosophical method currently treated as a panacea for philosophical ills, with consequent privileges for its home branch. Once we consider other branches of philosophy, we notice much more philosophizing whose primary subject matter is not conceptual.

Biology and physics are not studies of thought. In their most theoretical reaches, they merge into the philosophy of biology and the philosophy of physics. Why then should philosophers of biology and philosophers of physics study only thought? Although they sometimes study what biologists' and physicists' concepts are or should be, sometimes they study what those concepts are concepts of, in an abstract and general manner. If the conceptual turn is incompatible with regarding such activities as legitimately philosophical, why take the conceptual turn?

There is a more central example. Much contemporary metaphysics is not primarily concerned with thought or language at all. Its goal

is to discover what fundamental kinds of things there are and what properties and relations they have, not to study the structure of our thought about them – perhaps we have no thought about them until it is initiated by metaphysicians. Contemporary metaphysics studies substances and essences, universals and particulars, space and time, possibility and necessity. Although nominalist or conceptualist reductions of all these matters have been attempted, such theories have no methodological priority and generally turn out to do scant justice to what they attempt to reduce.

The usual stories about the history of twentieth-century philosophy fail to fit much of the liveliest, exactest, and most creative achievements of the final third of that century: the revival of metaphysical theorizing, realist in spirit, often speculative, sometimes commonsensical, associated with Saul Kripke, David Lewis, Kit Fine, Peter van Inwagen, David Armstrong and many others: work that has, to cite just one example, made it anachronistic to dismiss essentialism as anachronistic.<sup>6</sup> On the traditional grand narrative schemes in the history of philosophy, this activity must be a throwback to pre-Kantian metaphysics: it ought not to be happening – but it is. Many of those who practice it happily acknowledge its continuity with traditional metaphysics; appeals to the authority of Kant, or Wittgenstein, or history, ring hollow, for they are unbacked by any argument that has withstood the test of recent time.

One might try to see in contemporary metaphysics a Quinean breakdown of divisions between philosophy and the natural sciences. But if it is metaphysics naturalized, then so is the metaphysics of Aristotle, Descartes, and Leibniz. Armchair argument retains a central role, as do the modal notions of metaphysical possibility and necessity. Although empirical knowledge constrains the attribution of essential properties, results are more often reached through a subtle interplay of logic and the imagination. The crucial experiments are thought experiments.

Might the contrast between the new-old metaphysics and the conceptual turn be less stark than it appears to be? Contemporary metaphysicians firmly resist attempts to reconstrue their enterprise as

<sup>6</sup> On essentialism see, for example, Kripke (1980), French, Uehling, and Wettstein (1986), Fine (1994, 1995) and Wiggins (2001). For a good statement of the outlook of contemporary metaphysics see Zimmerman (2004).

the analysis of thought – unlike Sir Peter Strawson, who defined his “descriptive metaphysics” as “content to describe the actual structure of our thought about the world” (1959: 9). But can one reflect on concepts without reflecting on reality itself? For the aboutness of thought and talk is their very point. This idea has been emphasized by David Wiggins, Dummett’s successor and my predecessor in the Wykeham Chair, and author of some of the most distinguished essentialist metaphysics, in which considerations of logic and biology harmoniously combine. Wiggins (2001: 12) writes: “Let us forget once and for all the very idea of some knowledge of language or meaning that is not knowledge of the world itself.”

Wiggins is not just stating the obvious, that language and meaning are part of the world because everything is part of the world. Rather, his point is that in defining words – natural kind terms, for instance – we must point at real specimens. What there is determines what there is for us to mean. In knowing what we mean, we know something about what there is. That prompts the question how far the analysis of thought or language can be pursued autonomously with any kind of methodological priority.

Dummett claimed not that the traditional questions of metaphysics cannot be answered but that the way to answer them is by the analysis of thought and language. For example, in order to determine whether there are numbers, one must determine whether number words such as “7” function semantically like proper names in the context of sentences uttered in mathematical discourse. But what is it so to function? Although devil words such as “Satan” appear to function semantically like proper names in the context of sentences uttered in devil-worshipping discourse, one should not jump to the conclusion that there are devils. However enthusiastically devil-worshippers use “Satan” as though it referred to something, that does not make it refer to something. Although empty names *appear* to function semantically like referring names in the context of sentences uttered by those who believe the names to refer, the appearances are deceptive. “Satan” refers to something if and only if some sentence with “Satan” in subject position (such as “Satan is self-identical”) expresses a truth, but the analysis of thought and language is not the best way to discover whether any such sentence does indeed express a truth. Of course, what goes for “Satan” may not go for “7.” According to some neo-logicists, “7 exists” is an analytic truth (what

Ayer might have called a formal consequence of definitions), which “Satan exists” does not even purport to be. Such a claim needs the backing of an appropriate theory of analyticity.

After this preliminary sketch, it is time to get down to detailed work. The next three chapters examine different forms of the linguistic or conceptual turn. Chapter 2 uses a case study to consider in a microcosm the idea that philosophers’ questions are implicitly about language or thought when they are not explicitly so. Chapters 3 and 4 assess a wide range of versions of the idea that the armchair methodology of philosophy is grounded in the analytic or conceptual status of a core of philosophical truths, which need not be *about* language or thought, even implicitly. In each case the upshot is negative. Although philosophers have more reason than physicists to consider matters of language or thought, philosophy is in no deep sense a linguistic or conceptual inquiry, any more than physics is. But it does not follow that experiment is an appropriate primary method for philosophy. Similar arguments suggest that mathematics is in no deep sense a linguistic or conceptual inquiry, yet experiment is not an appropriate primary method for mathematics. The second half of the book develops an alternative conception of philosophy, on which a largely armchair methodology remains defensible, as it does for mathematics.

From this perspective and that of many contemporary philosophers, the conceptual turn and *a fortiori* the linguistic turn look like wrong turnings. It is pointless to deny that such philosophers are “analytic,” for that term is customarily applied to a broad, loose tradition held together by an intricate network of causal ties of influence and communication, not by shared essential properties of doctrine or method: what do Frege, Russell, Moore, Wittgenstein, Carnap, Ayer, Quine, Austin, Strawson, Davidson, Rawls, Williams, Anscombe, Geach, Armstrong, Smart, Fodor, Dummett, Wiggins, Marcus, Hintikka, Kaplan, Lewis, Kripke, Fine, van Inwagen and Stalnaker all have in common to distinguish them from all the non-analytic philosophers? Many who regard the linguistic and conceptual turns as serious mistakes have ties of influence and communication that put them squarely within that tradition. “Analytic philosophy” is a phrase in a living language; the attempt to stipulate a sense for it that excludes many of the philosophers just listed will achieve nothing but brief terminological confusion.

Historians of philosophy on the grand scale may be too Whiggish or Hegelian to regard the linguistic or conceptual turn as merely a false turning from which philosophy is withdrawing now that it recognizes its mistake. We are supposed to go forward from it, not back. At the very least, we should learn from our mistakes, if only not to repeat them. But if the conceptual turn was a mistake, it was not a simple blunder; it went too deep for that. A new narrative structure is needed for the history of philosophy since 1960; it is clear only in the roughest outline what it should be.

# Taking Philosophical Questions at Face Value

---

How often are philosophical questions implicitly about thought or language when they are not explicitly so? As a case study, I will take a question closely related to the problem of vagueness, because it looks like a paradigm of a philosophical question that is implicitly but not explicitly about thought and language. For vagueness is generally conceived as a feature of our thought and talk about the world, not of the world itself. Admittedly, some philosophers find tempting the idea of mind-independently vague objects, such as Mount Everest, vague in their spatiotemporal boundaries and mereological composition, if not in their identity. That kind of vagueness is not my concern here. I will consider an example of a quite standard type, involving a vague predicate.<sup>1</sup> Yet the reconstrual of the question as implicitly about thought or language turns out to be a mistake. If it is a mistake here, in such favorable conditions, it is a mistake far more widely.

Suppose that there was once plenty of water on the planet Mars; it was clearly not dry. Ages passed, and very gradually the water evaporated. Now Mars is clearly dry. No moment was clearly the first on which it was dry or the last on which it was not. For a long intermediate period it was neither clearly dry nor clearly not dry. Counting the water molecules would not have enabled us to determine whether

<sup>1</sup> On vagueness in general see, for a start, Graff and Williamson (2002), Keefe (2000), Keefe and Smith (1997), and Williamson (1994a). On vague objects see Williamson (2003b) and references therein.

it was dry; other measures would have been equally inconclusive. We have no idea of any investigative procedure that would have resolved the issue. It was a borderline case. No urgent practical purpose compels us to ask whether Mars was dry then, but only a limited proportion of thought and talk in any human society is driven by urgent practical purposes. We should like to know the history of Mars. When necessary, we can always use words other than “dry.” Nevertheless, we reflect on the difficulty of classifying Mars as dry or as not dry at those intermediate times, even given exact measurements. We may wonder whether it was either. We ask ourselves:

Was Mars always either dry or not dry?

Henceforth I will refer to that as *the original question*. More precisely, I will use that phrase to designate that interrogative sentence, as used in that context (the word “question” can also be applied to what interrogative sentences express rather than the sentences themselves).

The original question is at least proto-philosophical in character. It is prompted by a difficulty both hard to identify and hard to avoid that we encounter in applying the distinctions in our repertoire. It hints at a serious threat to the validity of our most fundamental forms of deductive reasoning. Philosophers disagree about its answer, on philosophical grounds explored below. A philosophical account of vagueness that does not tell us how to answer the original question is thereby incomplete. Without an agreed definition of “philosophy,” we can hardly expect to *prove* that the original question or any other is a philosophical question; but when we discuss its answer, we find ourselves invoking recognizably philosophical considerations. More simply, I’m a philosopher, I find the original question interesting, although I think I know the answer, and I have no idea where one should go for an answer to it, if not to philosophy (which includes logic). But before we worry about the answer, let us examine the original question itself.

The question queries just the supposition that Mars was always either dry or not dry, which we can formalize as a theorem of classical logic,  $\forall t (\text{Dry}(m, t) \vee \neg \text{Dry}(m, t))$ .<sup>2</sup> In words: for every time  $t$ ,

<sup>2</sup> Classical logic is the standard logic of expressions such as “every,” “either . . . or . . .” and “not” on the assumption that there is a mutually exclusive, jointly exhaustive dichotomy of sentences into the true and the false.

either Mars was dry at  $t$  or Mars was not dry at  $t$ . The question is composed of expressions that are not distinctively philosophical in character: “Mars,” “always,” “either . . . or . . . ,” “not,” “was,” and “dry.” All of them occur in a recognizably unphilosophical question such as “Was Mars always either uninhabited or not dry?,” which someone might ask on judging that Mars is both uninhabited and dry and wondering whether there is a connection. Although philosophical issues can be raised *about* the words in both questions, it does not follow that merely in using those words one is in any way engaging in philosophy. One difference between the two questions is that it is not obviously futile to try to argue from the armchair that Mars was always either dry or not dry, whereas it is obviously futile to try to argue from the armchair that Mars was always either uninhabited or not dry.

The original question does not itself *ask* whether it is metaphysically necessary, or knowable *a priori*, or analytic, or logically true that Mars was always either dry or not dry. It simply asks whether Mars always *was* either dry or not dry. Expressions such as “metaphysically necessary,” “knowable *a priori*,” “analytic,” and “logically true” do not occur in the original question; one can understand it without understanding any such philosophical terms of art. This is of course neither to deny nor to assert that it *is* metaphysically necessary, or knowable *a priori*, or analytic, or logically true that Mars was always either dry or not dry. For all that has been said, the proposition may be any combination of those things. But that is not what the original question asks.

In other circumstances, we could have answered the original question on philosophically uninteresting grounds. For instance, if there had never been liquid on Mars, then it would always have been dry, and therefore either dry or not dry. In order to pose a question which could not possibly be answered in that boring way, someone who already grasped one of those philosophically distinctive concepts might ask whether it is metaphysically necessary, or knowable *a priori*, or analytic, or logically true that Mars was always either dry or not dry. The meaningfulness of the philosophical jargon might then fall under various kinds of suspicion, which would extend to the question in which it occurred. But the original question itself cannot be correctly answered in the boring way with respect to the originally envisaged circumstances. Its philosophical interest, however contingent, is actual.

We could generalize the original question in various ways. We might ask whether *everything* is always either dry or not dry. Then we might notice that discussing that question is quite similar to discussing whether everything is either old or not old, and so on. We might, therefore, ask whether for every property everything either has it or lacks it. The coherence of such generalizing over properties might itself fall under various kinds of suspicion, which would extend to the question in which it occurred. Someone might even doubt whether there is such a property as dryness. But the original question itself does not attempt such generality. That it has the same kind of philosophical interest as many other questions does not imply that it has itself no philosophical interest. If that interest is obscured by problematic features of the apparatus with which we try to generalize it, we can refrain from generalizing it, and stick with the original question. In order not to be distracted by extraneous issues that arise from the apparatus of generalization, not from the original question, we do best to stick with the original question in its concrete form.<sup>3</sup> We can still help ourselves not to be distracted by unimportant features of the question, if we remember that there are many other questions of a similar form.

What is the original question about? “About” is not a precise term. On the most straightforward interpretation, a sentence in a context is about whatever its constituents refer to in that context. Thus, taken at face value, the original question is about the planet Mars, the referent of “Mars” in this context; perhaps it is also about dryness, the referent of “dry,” and the referents of other constituents too. Since the original question contains no metalinguistic expressions, it is not about the name “Mars” or the adjective “dry.” Evidently, the original question is not explicitly about words.

Is the original question implicitly about language? Someone might claim so on the grounds that it is equivalent to questions that are explicitly about language, such as these:

Is the sentence “Mars was always either dry or not dry” true? (Does it express a truth as used in this context?)

Did Mars always belong either to the extension of the word “dry” or to the anti-extension of “dry” (as the word “dry” is used in this context)?

<sup>3</sup> See also Quine (1970: 11).

But parallel reasoning would lead to the conclusion that the unphilosophical question “Was Mars always either uninhabited or not dry?” is also implicitly about language, since it is equivalent to these questions:

Is the sentence “Mars was always either uninhabited or not dry” true?  
(Does it express a truth as used in this context?)

Did Mars always belong either to the extension of the word “uninhabited” or to the anti-extension of “dry” (as the word “dry” is used in this context)?

Indeed, we could make parallel arguments for all everyday and scientific questions. Since they are not all about language in any distinctive sense, the reasoning does not show that the original question was about language in any distinctive sense. Even if the equivalences did show that the original question was in some sense implicitly about language, they could be read in both directions: they would also show that the explicitly metalinguistic questions were in an equally good sense implicitly not about language.

The equivalences between the questions are in any case uncontroversial only if the corresponding disquotational biconditionals are:

- (T1) “Mars was always either dry or not dry” is true if and only if Mars was always either dry or not dry.
- (T2a) For any time  $t$ , Mars belongs to the extension of “dry” at  $t$  if and only if Mars is dry at  $t$ .
- (T2b) For any time  $t$ , Mars belongs to the anti-extension of “dry” at  $t$  if and only if Mars is not dry at  $t$ .

On the face of it, these biconditionals express at best contingent truths. For perhaps the word “dry” could have meant *wet*, in which case Mars would have belonged to the extension of “dry” when wet and to the anti-extension of “dry” when dry: for *we* use the word “dry” to mean *dry* even when we are talking *about* circumstances in which it would have meant something else, because we are not talking *in* those circumstances. If so, T2a and T2b do not express necessary truths. Similarly, perhaps the sentence “Mars was always either dry or not dry” could have failed to express a truth even though Mars

was always either dry or not dry, since “always” could have meant *never*. On this reading, T1 does not express a necessary truth. We should not assume that a useful notion of aboutness would transfer across merely contingent biconditionals. Perhaps we can instead interpret T1, T2a, and T2b as expressing necessary truths by individuating linguistic expressions so that their semantic properties are essential to them; whether that requires treating the quoted expressions as necessary existents is a delicate matter. In any case, some theorists of vagueness have denied even the actual truth of biconditionals such as T1, T2a, and T2b; they might respond to the original question in one way and to the explicitly metalinguistic questions in another.<sup>4</sup> Thus the questions are not pragmatically, dialectically or methodologically equivalent within the context of debates on vagueness. For present purposes, we need not resolve the status of the disquotational biconditionals, because we have already seen that the sense in which they make the original question implicitly about words is too indiscriminate to be useful.

We can argue more directly that the original question is not implicitly about the word “dry” by appeal to a translation test. For consider the translation of the original question into another language, such as Serbian:

Da li je Mars uvek bio suv ili nije bio suv?

The Serbian translation is not implicitly about the English word “dry.” But since the questions in the two languages mean the same, what they are implicitly about (in the same context) should also be the same. Therefore, the original question is not implicitly about the word “dry.” By similar reasoning, it is not about any word of English or any other language. Of course, given the informality of the notion of implicit aboutness, the argument is not fully rigorous. Nevertheless, the translation test emphasizes how far one would have to water down the notion of reference in order to reach a notion of implicit aboutness on which the original question would be implicitly about a word.

<sup>4</sup> A recent example of a supervaluationist rejecting such disquotational equivalences for borderline cases is Keefe (2000: 213–20). For further discussion see Williamson (1994a: 162–4) and McGee and McLaughlin (2000).

The translation test does not show that the original question is not implicitly about a *concept*, something like the meaning of a word rather than the word itself, for the English word “dry” and its Serbian synonym “suv” both express the concept *dry*. But what basis is there for the claim that the original question is implicitly about the concept *dry*? We might argue that the original question is in some sense equivalent to a metaconceptual question:

Did Mars always belong either to the extension of the concept *dry* or to the anti-extension of *dry*?

For we might apply the notions of extension and anti-extension to concepts by means of biconditionals similar to T2a and T2b respectively:

- (TC2a) For any time  $t$ , Mars belongs to the extension of *dry* at  $t$  if and only if Mars is dry at  $t$ .
- (TC2b) For any time  $t$ , Mars belongs to the anti-extension of *dry* at  $t$  if and only if Mars is not dry at  $t$ .

TC2a and TC2b can express necessary truths more easily than T2a and T2b can, for the apparently contingent relation between words and their meanings has no straightforward analogue for concepts. Concepts are individuated semantically: rather than merely having meanings, they *are* meanings, or something like them.<sup>5</sup> Nevertheless, the argument that the original question is implicitly about the concept *dry* in virtue of being equivalent to the metaconceptual question wildly overgeneralizes, just like the argument that the original

<sup>5</sup> Even if a word retains its linguistic meaning, its reference may shift with the context of utterance (“I,” “now,” “here”). If “dry” undergoes such contextual shifts, T2a and T2b may fail when interpreted as generalizations about utterances of “dry” in contexts other than the theorist’s own. It might be argued that concepts can also undergo contextual shifts in reference: you use the concept *I* to refer (in thought) to yourself but I use the same concept to refer to myself; at noon we use the concept *now* to think of noon but at midnight we use the same concept to refer to midnight; at the North Pole we use the concept *here* to refer to the North Pole but at the South Pole we use the same concept to refer to the South Pole. If so, TC2a and TC2b may also fail when interpreted as generalizations about uses of the concept *dry* in contexts other than the theorist’s own.

question is implicitly about the word “dry” in virtue of being equivalent to the metalinguistic question. For parallel reasoning would lead to the conclusion that the unphilosophical question “Was Mars always either uninhabited or not dry?” is implicitly about the concept *dry*, and likewise for any other unphilosophical question. Since those questions are not about concepts in any distinctive sense, the original reasoning does not show that the original question is about concepts in any distinctive sense. Even if the equivalences did show that the original question was in some sense implicitly about thought, they can be read in both directions: they would also equally show that the explicitly metaconceptual questions were in an equally good sense implicitly not about thought.

A Fregean might argue: the original question is *explicitly* about the concept *dry*, because it contains the predicate “... is dry” (in the past tense), which refers to the concept *dry*. In that sense, the question “Was Mars always either uninhabited or not dry?” would also be explicitly about the concept *dry*. However, the Fregean is not using the word “concept” with its contemporary meaning, on which concepts are something like mental or semantic representations, closer to the realm of sense than to that of reference. The Fregean referent of a predicate (a Fregean concept) is simply the function that maps everything to which the predicate applies to the true and everything else to the false: it could be treated as the extension of the predicate, except that in Fregean terms it is a function rather than an object. If the predicate refers to the property of dryness or to the set of dry things, then the original question is about the property of dryness or the set of dry things, but that has no tendency to show that it is about thought. Similarly, the Fregean claim has no tendency to show that the question is about thought, for the Fregean concept is in the realm of reference, not in the realm of thought. Like the property and the set, it is no sense but something to which a sense may determine reference. Since it is no sense, it is no constituent of a thought, on the Fregean view, nor is it a concept in the current sense of “concept.”

Thought and talk are not always about thought or talk. To judge by its overt compositional structure, the original question in particular is not about thought or talk. It is no metalinguistic or metaconceptual question. We have seen no reason to regard its overt structure as at all misleading in that respect. Our provisional conclusion must therefore be that the original question, although at least proto-

philosophical, is not about thought or language in any distinctive sense. It does not support the linguistic or conceptual turn, interpreted as a conception of the subject matter of philosophy.

## 2

If the original question, read literally, had too obvious an answer, either positive or negative, that would give us reason to suspect that someone who uttered it had some other meaning in mind, to which the overt compositional structure of the question might be a poor guide. But competent speakers of English may find themselves quite unsure how to answer the question, read literally, so we have no such reason for interpreting it non-literally.

It is useful to look at some proposals and arguments from the vagueness debate, for two reasons. First, they show why the original question is hard, when taken at face value. Second, they show how semantic considerations play a central role in the attempt to answer it, even though it is not itself a semantic question.

The most straightforward reason for answering the original question positively is that “Mars was always either dry or not dry” is a logical truth, a generalization over instances of the law of excluded middle ( $A \vee \neg A$ , “It is either so or not so”) for various times. In my view, that reasoning is sound. However, many think otherwise. They deny the validity of excluded middle for vague terms such as “dry.”

The simplest way of opposing the law of excluded middle is to deny outright when Mars is a borderline case that it is either dry or not dry, and therefore to answer the original question in the negative. For instance, someone may hold that Mars was either dry or not dry at time  $t$  only if one can know (perhaps later) whether it was dry at  $t$ , given optimal conditions for answering the question (and no difference in the history of Mars): since one cannot know, even under such conditions, whether it is dry when the case is borderline, it is not either dry or not dry. One difficulty for this negative response to the original question is that it seems to imply that in a borderline case Mars is neither dry nor not dry: in other words, both not dry and not not dry. That is a contradiction, for “not not dry” is the negation of “not dry.”

Intuitionistic logic provides a subtler way to reject the law of excluded middle without denying any one of its instances. Intuitionists ground logic in states of increasing but incomplete information, rather than a once-for-all dichotomy of truth and falsity. They deny that anything can be both proved and refuted, but they do not assert that everything can be either proved or refuted. For intuitionists, the denial of an instance of excluded middle ( $\neg(A \vee \neg A)$ , “It is not either so or not so”) entails a contradiction ( $\neg A \& \neg\neg A$ , “It is both not so and not not so”), just as it does in classical logic, and contradictions are as bad for them as for anyone else. Thus they cannot assert that Mars was once not either dry or not dry ( $\exists t \neg(Dry(m, t) \vee \neg Dry(m, t))$ ), for that would imply that a contradiction once obtained ( $\exists t (\neg Dry(m, t) \& \neg\neg Dry(m, t))$ , “Mars was once both not dry and not not dry”), which is intuitionistically inconsistent. However, although intuitionists insist that proving an existential claim in principle involves proving at least one instance of it, they allow that disproving a universal claim need not in principle involve disproving at least one instance of it. The claim that something lacks a property is intuitionistically stronger than the claim that not everything has that property. Thus one might assert that Mars was not always either dry or not dry ( $\neg\forall t (Dry(m, t) \vee \neg Dry(m, t))$ ), on the general grounds that there is no adequate procedure for sorting all the times into the two categories, without thereby committing oneself to the inconsistent existential assertion that it was once not either dry or not dry. Hilary Putnam once proposed the application of intuitionistic logic to the problem of vagueness for closely related reasons.<sup>6</sup> Thus one might use intuitionistic logic to answer the original question in the negative.

On closer inspection, this strategy looks less promising. For a paradigm borderline case is the worst case for the law of excluded middle (for a term such as “dry” for which threats to the law other than from vagueness are irrelevant), in the sense that both proponents and opponents of the law can agree that it holds in a paradigm borderline case only if it holds universally. In symbols, if Mars was a paradigm borderline case at time  $\tau$ :  $(Dry(m, \tau) \vee \neg Dry(m, \tau)) \rightarrow$

<sup>6</sup> For intuitionist logic in general see Dummett (1977). For its application to the problem of vagueness see Graff and Williamson (2002: 473–506) and Chambers (1998).

$\forall t (Dry(m, t) \vee \neg Dry(m, t))$  (“If Mars was either dry or not dry at time  $\tau$ , then Mars was always either dry or not dry”). But on this approach the law does not hold always hold in these cases ( $\neg \forall t (Dry(m, t) \vee \neg Dry(m, t))$ ), “Mars was not always either dry or not dry”), from which intuitionistic logic allows us to deduce that it does not hold in the paradigm borderline case ( $\neg (Dry(m, \tau) \vee \neg Dry(m, \tau))$ ), “Mars was not either dry or not dry at  $\tau$ ”), which is a denial of a particular instance of the law, and therefore intuitionistically inconsistent (it entails  $\neg Dry(m, \tau) \& \neg \neg Dry(m, \tau)$ , “Mars was both not dry and not not dry at  $\tau$ ”). Thus the intuitionistic denial of the universal generalization of excluded middle for a vague predicate forces one to deny that it has such paradigm borderline cases. The latter denial is hard to reconcile with experience: after all, the notion of a borderline case is usually explained by examples.

The problems for the intuitionistic approach do not end there. One can show that the denial of the conjunction of any finite number of instances of the law of excluded middle is intuitionistically inconsistent.<sup>7</sup> The denial of the universal generalization of the law over a finite domain is therefore intuitionistically false too. If time is infinitely divisible, the formula  $\forall t (Dry(m, t) \vee \neg Dry(m, t))$  generalizes the law over an infinite domain of moments of time, and its denial is intuitionistically consistent, but the possibility of infinitely divisible time is not crucial to the phenomena of vagueness. We could just as well have asked the original question about a long finite series of moments at one-second intervals; it would have been equally problematic. The classical sorites paradox depends on just such a finite series: a heap of sand consists of only finitely many grains, but when they are carefully removed one by one, we have no idea how to answer the question “When did there cease to be a heap?” To deny that Mars was dry or not dry at each moment in the finite series is intuitionistically inconsistent. Thus intuitionistic logic provides a poor basis for a negative answer to the original question.

Other theorists of vagueness refuse to answer the original question either positively or negatively. They refuse to assert that Mars was always either dry or not dry; they also refuse to assert that it was not always either dry or not dry.

<sup>7</sup> One proves by mathematical induction on  $n$  that if  $\mathbf{An}$  is the conjunction of  $n$  instances of excluded middle then  $\neg \mathbf{An}$  is intuitionistically inconsistent.

A simple version of this approach classifies vague sentences (relative to contexts) as true (T), false (F) or indefinite (I); borderline sentences are classified as indefinite. The generalized truth-tables of a three-valued logic are used to calculate which of these values to assign to a complex sentence in terms of the values assigned to its constituent sentences. The negation of  $A$ ,  $\neg A$ , is true if  $A$  is false, false if  $A$  is true and indefinite if  $A$  is indefinite:

$A$	$\neg A$
T	F
I	I
F	T

A conjunction  $A \ \& \ B$  (“A and B”) is true if every conjunct is true; it is false if some conjunct is false; otherwise it is indefinite. A disjunction  $A \vee B$  (“Either A or B”) is true if some disjunct is true; it is false if every disjunct is false; otherwise it is indefinite:

$A$	$B$	$A \ \& \ B$	$A \vee B$
T	T	T	T
T	I	I	T
T	F	F	T
I	T	I	T
I	I	I	I
I	F	F	I
F	T	F	T
F	I	F	I
F	F	F	F

A universal generalization is treated as if it were the conjunction of its instances, one for each member of the domain: it is true if every instance is true, false if some instance is false, and otherwise indefinite. An existential generalization is treated as if it were the disjunction of the instances: it is true if some instance is true, false if every instance is false, and otherwise indefinite. The three-valued tables generalize the familiar two-valued ones in the sense that one recovers the latter by deleting all lines with “I.”

Let us apply this three-valued approach to the original question. If Mars is definitely dry or definitely not dry at  $t$  (the time denoted by  $t$ ), then  $\text{Dry}(m, t)$  is true or false, so the instance of excluded middle

$\text{Dry}(m, t) \vee \neg \text{Dry}(m, t)$  is true. But if Mars is neither definitely dry nor definitely not dry at  $t$ , then  $\text{Dry}(m, t)$  is indefinite, so  $\neg \text{Dry}(m, t)$  is indefinite too by the table for negation, so  $\text{Dry}(m, t) \vee \neg \text{Dry}(m, t)$  is classified as indefinite by the table for disjunction. Since Mars was once a borderline case, the universal generalization  $\forall t (\text{Dry}(m, t) \vee \neg \text{Dry}(m, t))$  has a mixture of true and indefinite instances; hence it is classified as indefinite. Therefore its negation  $\neg \forall t (\text{Dry}(m, t) \vee \neg \text{Dry}(m, t))$  is also indefinite. Thus three-valued theoreticians who wish to assert only truths neither assert  $\forall t (\text{Dry}(m, t) \vee \neg \text{Dry}(m, t))$  nor assert  $\neg \forall t (\text{Dry}(m, t) \vee \neg \text{Dry}(m, t))$ . They answer the original question neither positively nor negatively.

Three-valued logic replaces the classical dichotomy of truth and falsity by a three-way classification. Fuzzy logic goes further, replacing it by a continuum of degrees of truth between perfect truth and perfect falsity. According to proponents of fuzzy logic, vagueness should be understood in terms of this continuum of degrees of truth. For example, “It is dark” may increase continuously in degree of truth as it gradually becomes dark. On the simplest version of the approach, degrees of truth are identified with real numbers in the interval from 0 to 1, with 1 as perfect truth and 0 as perfect falsity. The semantics of fuzzy logic provides rules for calculating the degree of truth of a complex sentence in terms of the degrees of truth of its constituent sentences. For example, the degrees of truth of a sentence and of its negation sum to exactly 1; the degree of truth of a disjunction is the maximum of the degrees of truth of its disjuncts; the degree of truth of a conjunction is the minimum of the degrees of truth of its conjuncts. For fuzzy logic, although the three-valued tables above are too coarse-grained to give complete information, they still give correct results if one classifies every sentence with an intermediate degree of truth, less than the maximum and more than the minimum, as indefinite.<sup>8</sup> Thus the same reasoning as before shows that fuzzy

<sup>8</sup> This point does not generalize to the semantics of conditionals in fuzzy logic, given the popular rule that if the consequent is lower than the antecedent in degree of truth then the degree of truth of the conditional falls short of 1 by the amount by which the consequent falls short of the antecedent in degree of truth; otherwise the degree of truth of the conditional is 1. Hence if  $A$  has a higher degree of truth than  $B$  but both are indefinite then  $A \rightarrow B$  is indefinite while  $B \rightarrow A$  is perfectly true. Thus the information that the antecedent and consequent are indefinite does not determine whether the conditional is indefinite.

logicians should answer the original question neither positively nor negatively.

Although three-valued and fuzzy logicians reject both the answer “Yes” and the answer “No” to the original question, they do not reject the question itself. What they reject is the restriction of possible answers to “Yes” and “No.” They require a third answer, “Indefinite,” when the queried sentence takes the value I. More formally, consider the three-valued table for the sentence operator  $\Delta$ , read as “definitely” or “it is definite that”:

A	$\Delta A$
T	T
I	F
F	F

Even for fuzzy logicians this table constitutes a complete semantics for  $\Delta$ , since the only output values are T and F, which determine unique degrees of truth (1 and 0). A formula of the form  $\neg\Delta A \ \& \ \neg\Delta\neg A$  (“It is neither definitely so nor definitely not so”) characterizes a borderline case, for it is true if A is indefinite and false otherwise. In response to the question A?, answering “Yes” is tantamount to asserting A, answering “No” is tantamount to asserting  $\neg A$ , and answering “Indefinite” is tantamount to asserting  $\neg\Delta A \ \& \ \neg\Delta\neg A$ . On the three-valued and fuzzy tables, exactly one of these three answers is true in any given case; in particular, the correct answer to the original question is “Indefinite.”

On the three-valued and fuzzy approaches, to answer “Indefinite” to the question “Is Mars dry?” is to say something about Mars, just as it is if one answers “Yes” or “No.” It is not a metalinguistic response. For  $\Delta$  is no more a metalinguistic operator than  $\neg$  is. They have the same kind of semantics, given by a many-valued truth-table. Just as the negation  $\neg A$  is about whatever A is about, so are  $\Delta A$  and  $\neg\Delta A \ \& \ \neg\Delta\neg A$ . Thus the answer “Indefinite” to the original question involves no semantic ascent to a metalinguistic or metaconceptual level. It remains at the level of discourse about Mars.

The three-valued and fuzzy approaches have many suspect features. For instance, they treat any sentence of the form  $\Delta A$  as perfectly precise, because it always counts as true or false, never as indefinite, whatever the status of A; thus  $\Delta\Delta A \ \vee \ \Delta\neg\Delta A$  (“It is definite whether it is definitely so”) is always true. This result does not fit the intended

interpretation of  $\Delta$ . For “Mars is definitely wet” is not perfectly precise. Just as no moment is clearly the last on which Mars was wet or the first on which it was not, so no moment is clearly the last on which it was definitely wet or the first on which it was not definitely wet. Just as it is sometimes unclear whether Mars is wet, so it is sometimes unclear whether it is definitely wet. This is one form of the notorious problem of higher-order vagueness: in other words, there are borderline cases of borderline cases, and borderline cases of borderline cases of borderline cases, and so on. The problem has never received an adequate treatment within the framework of three-valued or fuzzy logic; that it could is far from obvious.<sup>9</sup>

Some philosophers, often under the influence of the later Wittgenstein, deny the relevance of formal semantic theories to vague natural languages. They regard the attempt to give a systematic statement of the truth conditions of English sentences in terms of the meanings of their constituents as vain. For them, the formalization of “Mars was always either dry or not dry” as  $\forall t (Dry(m, t) \vee \neg Dry(m, t))$  is already a mistake. This attitude suggests a premature and slightly facile pessimism. No doubt formal semantics has not described any natural language with perfect accuracy; what has not been made plausible is that it provides no deep insights into natural languages. In particular, it has not been made plausible that the main semantic effects of vagueness are not susceptible to systematic formal analysis. In any case, for present purposes the claim that there can be no systematic theory of vagueness is just one more theory of vagueness, although – unless it is self-refuting – not a systematic one; it does not even answer the original question. Even if that theory were true, the other theories of vagueness, however false, would still exist, and would still have been accepted by some intelligent and linguistically competent speakers.

This is no place to resolve the debate between opposing theories of vagueness. The present point is just that different theories support contrary answers to the original question. All these theories have their believers. Any answer to the original question, positive, negative, or indefinite, is contentious. Of course, if everyone found their own answer obvious, but different people found different answers obvious, then we might suspect that they were interpreting the question in

<sup>9</sup> See Graff and Williamson (2002: 279–351) on higher-order vagueness.

different ways, talking past each other. But that is not so: almost everyone who reflects on the original question finds it difficult and puzzling. Even when one has settled on an answer, one can see how intelligent and reasonable people could answer differently while understanding the meaning of the question in the same way. If it has an *obvious* answer, it is the answer “Yes” dictated by classical logic, but those of us who accept that answer can usually imagine or remember the frame of mind in which one is led to doubt it. Thus the original question, read literally, has no unproblematically obvious answer in any sense that would give us reason to suspect that someone who asked it had some other reading in mind.

Without recourse to non-literal readings, some theorists postulate ambiguity in the original question. For example, some three-valued logicians claim that “not” in English is ambiguous between the operators  $\neg$  (strong negation) and  $\neg\Delta$  (weak negation): although  $\neg A$  and  $\neg\Delta A$  have the same value if  $A$  is true or false,  $\neg\Delta A$  is true while  $\neg A$  is indefinite if  $A$  is indefinite. While  $A \vee \neg A$  (“It is so or not so”) can be indefinite,  $A \vee \neg\Delta A$  (“It is so or not definitely so”) is always true. On this view, the original question queries  $\forall t (\text{Dry}(m, t) \vee \neg\text{Dry}(m, t))$  on one reading,  $\forall t (\text{Dry}(m, t) \vee \neg\Delta\text{Dry}(m, t))$  on another; the latter is true (Mars was always either dry or not definitely dry) while the former is indefinite. Thus the correct answer to the original question depends on the reading of “not.” It is “Indefinite” if “not” is read as strong negation, “Yes” if “not” is read as weak negation. Although the three-valued logician’s reasoning here is undermined by higher-order vagueness, that is not the present issue.<sup>10</sup>

If “not” were ambiguous in the way indicated, it would still not follow that the dispute over the original question is merely verbal. For even when we agree to consider it under the reading of “not” as strong negation, which does not factorize in the manner of  $\neg\Delta$ , we still find theories of vagueness in dispute over the correct answer. We have merely explained our terms in order to formulate more clearly a difficult question about Mars.

Still, it might be suggested, the dispute between different theories of vagueness is verbal in the sense that their rival semantics characterize different possible languages or conceptual schemes: our choice of which of them to speak or think would be pragmatic, based on

<sup>10</sup> See Williamson (1994a: 193–5).

considerations of usefulness rather than of truth. Quine defended a similar view of alternative logics (1970: 81–6).

To make sense of the pragmatic view, suppose that the original vague atomic sentences are classifiable both according to the bivalent scheme as true or false and according to the trivalent scheme as definitely true, indefinite or definitely false, and that the truth-tables of each scheme define intelligible connectives, although the connective defined by a trivalent table should be distinguished from the similar-looking connective defined by the corresponding bivalent table. Definite truth implies truth, and definite falsity implies falsity, but indefiniteness does not discriminate between truth and falsity: although all borderline atomic sentences are indefinite, some are true and others false. As Mars dries, “Mars is dry” is first false and definitely false, then false but indefinite, then true but indefinite, and finally true and definitely true. However, this attempted reconciliation of the contrasting theories does justice to neither side. For trivalent logicians, once we know that a sentence is indefinite, there is no further question of its truth or falsity to which we do not know the answer: the category of the indefinite was introduced in order not to postulate such a mystery. Similarly, for fuzzy logicians, once we know the intermediate degree of truth of a sentence, there is no further question of its truth or falsity to which we do not know the answer: intermediate degrees of truth were introduced in order not to postulate such a mystery. In formal terms, trivalent and fuzzy logics are undoubtedly less convenient than bivalent logic; the justification for introducing them was supposed to be the inapplicability of the bivalent scheme to vague sentences. If a bivalent vague language is a genuinely possible option, then the trivalent and fuzzy accounts of vagueness are mistaken. Conversely, from a bivalent perspective, the trivalent and fuzzy semantics do not fix possible meanings for the connectives, because they do not determine truth conditions for the resultant complex sentences: for example, the trivalent table for  $\neg$  does not specify when  $\neg A$  is true in the bivalent sense. It would, therefore, be a fundamental misunderstanding of the issue at stake between theories of vagueness to conceive it as one of a pragmatic choice of language.

We already speak the language of the original question; we understand those words and how they are put together; we possess the concepts they express; we grasp what is being asked. That semantic

knowledge may be necessary if we are to know the answer to the original question.<sup>11</sup> It is not sufficient, for it does not by itself put one in a position to arbitrate between conflicting theories of vagueness. For each of those theories has been endorsed by some competent speakers of English who fully grasp the question.

Competent speakers may of course fail to reflect adequately on their competence. Although the proponents of conflicting theories of vagueness presumably have reflected on their competence, their reflections may have contained mistakes. Perhaps reflection of sufficient length and depth on one's competence would lead one to the correct answer to the original question. But the capacity for such more or less philosophical reflection is not a precondition of semantic competence. Philosophers should resist the professional temptation to require all speakers to be good at philosophy.

We can distinguish two levels of reflection, the logical and the metalogical. In response to the original question, logical reflection involves reasoning with terms of the kind in which the question is phrased; the aim is to reach a conclusion that answers the question. For example, one might conclude by classical logic that Mars was always either dry or not dry; one might conclude by fuzzy logic that it is indefinite whether it was always one or the other. The logical level is not purely mechanical. When the reasoning is complex, one needs skill to select from the many permissible applications of the rules one sequence that leads to an answer to the question. When the reasoning is informal, one needs good judgment to select only moves that really are permissible applications of the rules. But one is still thinking about whatever the question was about. One starts only at the metalogical level of reflection to think about the semantics of the logical connectives and other expressions one employed at the logical level. For example, at the metalogical level one may assert or deny

<sup>11</sup> Of course, monolingual speakers of another language may know whether Mars was always dry or not dry without ever hearing of the original question, which is an interrogative sentence of English; they use a synonymous sentence of their own language. They do not know whether the original English question has a positive answer. Someone may even know whether the original English question has a positive answer without understanding the question, because the knowledge can be passed along a chain of testimony; understanding of the original question is needed only at one end of the chain. These quibbles do not affect the argument.

that the sentence “Mars was always either dry or not dry” is a logical truth. The rules used at the logical level are articulated only at the metalogical level.

It must be possible to think logically without thinking metalogically, for otherwise by the same principle thinking metalogically would involve thinking met metalogically, and so *ad infinitum*: our thinking never goes all the way up such an infinite hierarchy. What can prompt ascent to the metalogical level are hard cases in which one feels unclear about the permissibility of a given move at the logical level. One’s mastery of the language and possession of concepts leave one quite uncertain how to go on. In the case of the original question, a salient line of classical reasoning leads to a positive answer: it persuades some competent speakers while leaving others unconvinced. Even to discuss the contentious reasoning we must semantically ascend. We cannot hope to resolve the dispute undogmatically if we never leave the lower level.

### 3

The argument so far has reached two conclusions at first sight hard to reconcile with each other. First, the original question is not about thought or language. Second, to answer it adequately one must assess rival theories of vagueness in thought and language. How can that way of reaching an answer be appropriate to the original question? We might, therefore, find ourselves tempted back to the idea that somehow the original question was surreptitiously about thought or language.

On further reflection, the combination of the two conclusions is less surprising. Many non-philosophical questions that are not about thought or language cannot be resolved without inquiry into thought or language. Suppose that a court of law must decide whether Smith killed Jones. The question is not who said or thought what. Nevertheless, the crucial arguments may be over whether to trust the witnesses’ testimony. How is what they say now related to what they think now or thought then? How is what they think now or thought then related to what actually happened? Are they lying or sincere? Are their memories confused or clear? Those are questions about their thought and speech. They hold the key to whether Smith killed Jones, even

though that question is not about thought about language.<sup>12</sup> Of course, the questions about the thought or talk are not about it in isolation from what it is thought or talk about: they are relevant because they concern the relation between the thought or talk and what it is about.

The court must decide the issue on the evidence before it. In a criminal case, does the evidence put it beyond reasonable doubt that Smith killed Jones? In a civil case, does the evidence make it more probable than not? If the court is really deciding a question about testimonial evidence, that is already a question about talk.<sup>13</sup> But the question about the evidence arises in virtue of its bearing on the primary question, whether Smith killed Jones. Indeed, the question about the evidence is exactly a question about its bearing on the primary question. So the point stands.

Historians are often in a similar position. They want to know what happened. The way to achieve that is largely by considering documents, linguistic accounts of what happened – not in isolation, but in relation to what they represent. Most obviously, historians want to know whether the documents accurately represent what happened, but to answer that question they must in turn ask about their provenance: who produced them, when and why? Thus the history of the events of primary interest requires a history of thought and talk about those events. Those histories typically overlap, for thought or talk about some part of a complex human event is often another part of the same complex event.

Something analogous occurs in the methodology of the natural sciences. We wish to know the value of some physical quantity. We must devise apparatus to measure it. We may find ourselves in disputes over the functioning of different devices. Although the primary

<sup>12</sup> The issue of Smith's intentions concerns his thoughts, but we may suppose that the question immediately at issue is whether Smith was even involved in Jones's death.

<sup>13</sup> Non-testimonial evidence may be taken to include non-linguistic items such as a bloodied knife; this is what lawyers call "real evidence." For an argument that all evidence in an epistemologically central sense of the term is propositional see Williamson (2000a: 194–200). For example, the evidence in this sense might include the proposition that the bloodied knife was found at the scene of the crime, but not the knife itself.

question was not about those measuring devices, we cannot answer it adequately without considering them. We need a theory about the relation between the value of the quantity and the representations of it we record when we use our instruments. The scientific investigation of the physical quantity widens to include the scientific investigation of its interaction with our experimental equipment. After all, our apparatus is part of the same natural world as the primary topic of our inquiry.

These analogies make it less surprising that when we try to answer the original question, which is not a question about thought or language, our main task is to adjudicate between rival theories of vague thought and language. A theory of vagueness validates some deduction that concludes with an answer to the original question. That deduction uses but does not mention vague thought or language. It is formulated at the logical level, like the original question itself, not at the metalogical level. But discursively to justify trusting that deduction, rather than one that reaches another conclusion by other rules, one must assess the rival theories of vagueness.

That theories of vagueness conflict in their answers to the original question shows that they are not confined to claims about thought and talk. Theories such as epistemicism and supervaluationism which employ classical logic have “Mars was always either dry or not dry” as a theorem, once they are formulated in a suitably expressive language. To reiterate, that theorem is not about thought or talk.

For the three-valued and fuzzy approaches, the matter is only slightly more complicated. Their proponents assert:

(C) It is indefinite whether Mars was always either dry or not dry.

On those approaches, C does not count as about thought or language. Strictly speaking, however, C does not follow from the three-valued or fuzzy theory of vagueness itself; for all the theory implies, there was never any liquid on Mars, in which case it would always have been either dry or not dry, even by three-valued or fuzzy standards, and so would not have been indefinite. The theory implies only a conditional theorem:

(P1) If it was once indefinite whether Mars was dry then it is indefinite whether Mars was always either dry or not dry.

Three-valued or fuzzy theorists can combine P1 with what they regard as an empirical truth about Mars:

(P2) It was once indefinite whether Mars was dry.

From P1 and P2 they use the rule of modus ponens (from “If P then Q” and “P” infer “Q”) to infer C, the answer to the original question. Although their theorem P1 does not answer the question by itself, it is no more about thought or language than C is. Their theories are just as committed as classical ones to making claims that are not about thought or language.

In principle, just as the considerations relevant to adjudicating the dispute between theories of vagueness are relevant to answering the original question, so too may they be relevant to answering a question asked with no philosophical intention, such as “Was Mars always either uninhabited or not dry?,” if it turns out to involve a borderline case. In practice, non-philosophers are often quite content to be told “It’s unclear,” without wondering exactly how that statement addresses the question asked; they simply drop the matter. For their purposes that may be the best thing to do. By contrast, philosophers persist; they want to know at least whether there is a right answer, even if nobody can know what it is. The difference lies not in the content of the original question but in the interests with which it is asked. Those interests can amount to a tissue of associated questions: for our original question as asked by a philosopher, the associated questions query other instances of the law of excluded middle. Given those interests, it is rational to persist with the original question, and not take an unexplained “It’s unclear” for an answer. But we should not underestimate the importance outside philosophy too – in science and even in politics – of sometimes persisting with a straight question, not allowing oneself to be fobbed off with the convenient claim that no practical purpose would be served by answering it. At other times, non-philosophers in effect assume without argument a particular treatment of vagueness (not always the same one), without realizing or caring that there are alternatives. The treatment may be good enough for their purposes, or not.

In this case study, our interest in giving a clear and critically reflective answer to a simple, non-technical, non-metalinguistic, non-metaconceptual question forced us to adjudicate between complex,

technical, metalinguistic, and metaconceptual theories. This phenomenon seems to have been overlooked by those who complain about the “arid” technical minuteness of much philosophy in the analytic tradition. A question may be easy to ask but hard to answer. Even if it is posed in dramatic and accessible terms, the reflections needed to select rationally between rival answers may be less dramatic and accessible. Such contrasts are commonplace in other disciplines; it would have been amazing if they had not occurred in philosophy. Impatience with the long haul of technical reflection is a form of shallowness, often thinly disguised by histrionic advocacy of depth. Serious philosophy is always likely to bore those with short attention-spans.<sup>14</sup>

Why should considerations about thought and language play so much more central a role in philosophy than in other disciplines, when the question explicitly under debate is not itself even implicitly about thought or language? The paradigms of philosophical questions are those that seem best addressed by armchair considerations less formal than mathematical proofs. The validity of such informal arguments depends on the structure of the natural language sentences in which they are at least partly formulated, or on the structure of the underlying thoughts. That structure is often hard to discern. We cannot just follow our instincts in reasoning; they are too often wrong (see Chapter 4 for details). In order to reason accurately in informal terms, we must focus on our reasoning as presented in thought or language, to double-check it, and the results are often controversial. Thus questions about the structure of thought and language become central to the debate, even when it is not primarily a debate about thought or language.

The rise of modern logic from Frege onwards has provided philosophers with conceptual instruments of unprecedented power and precision, enabling them to formulate hypotheses with more clarity and determine their consequences with more reliability than ever before. Russell’s theory of descriptions showed vividly how differences between the surface form of a sentence and its underlying semantic structure might mislead us as to its logical relations and thereby create philosophical illusions. The development of formal

<sup>14</sup> Popularization has its place, in philosophy as in physics, but should not be confused with the primary activity.

model-theory and truth-conditional semantics by Tarski and others has provided a rigorous framework for thinking about the validity of our inferences. These theoretical advances have enormous intellectual interest in their own right. They may have made it tempting to suppose that all philosophical problems are problems of language: but they do not really provide serious evidence for that conjecture.

To deny that all philosophical questions are about thought or language is not to deny the obvious, that many are. We have also seen how in practice the attempt to answer a question which is not about thought or language can largely consist in thinking about thought and language. Some contemporary metaphysicians appear to believe that they can safely ignore formal semantics and the philosophy of language because their interest is in a largely extra-mental reality. They resemble an astronomer who thinks he can safely ignore the physics of telescopes because his interest is in the extra-terrestrial universe. In delicate matters, his attitude makes him all the more likely to project features of his telescope confusedly onto the stars beyond. Similarly, the metaphysicians who most disdain language are the most likely to be its victims. Again, those who neglect logic in order to derive philosophical results from natural science make frequent logical errors in their derivations; their philosophical conclusions do not follow from their scientific premises. For example, some supposed tensions between folk theory and contemporary science depend on fallacies committed in the attempt to draw out the consequences of common sense beliefs.

Analytic philosophy at its best uses logical rigor and semantic sophistication to achieve a sharpness of philosophical vision unobtainable by other means. To sacrifice those gains would be to choose blurred vision. Fortunately, one can do more with good vision than look at eyes.

Many have been attracted to the idea that all philosophical problems are linguistic or conceptual through the question: if the method of philosophy is *a priori* reflection, how can it lead to substantive knowledge of the world? Those who find that question compelling may propose that it informs us of relations of ideas rather than matters of fact, or that its truths are analytic rather than synthetic, or that it presents rules of grammar disguised as descriptions, or that its aim is the analysis of thought or language. In short, on this view, philosophical truths are conceptual truths. We may suspect the pres-

ence of empiricist presuppositions in the background – or, as with Ayer, in the foreground. Not starting with such presuppositions, we should be open to the idea that thinking just as much as perceiving is a way of learning how things are. Even if one does not fully understand *how* thinking can provide new knowledge, the cases of logic and mathematics constitute overwhelming evidence that it does so. The case of the original question, which is philosophical yet queries a theorem of classical logic, shows that we cannot segregate logic from philosophy and claim that armchair thinking illuminates the former but not the latter. In particular, conceptions of logic and mathematics as (unlike philosophy) somehow trivial or non-substantial have not been vindicated by any clear explanation of the relevant sense of “trivial” or “non-substantial.” Whether a given formal system of logic or mathematics is consistent is itself a non-trivial question of logic or mathematics. We know from Gödel’s second incompleteness theorem that the consistency of most standard systems of elementary mathematics cannot be decided in equally elementary mathematics, unless the original system is already inconsistent. The next two chapters investigate in more depth the prospects for conceptual truth and its role in philosophy.

# Metaphysical Conceptions of Analyticity

---

“Philosophical questions are more conceptual in nature than those of other disciplines”: that can easily pass for a statement of the obvious.<sup>1</sup> Many philosophers consciously seek conceptual connections, conceptual necessities, conceptual truths, conceptual analyses. In effect, they present themselves as seeking far more general and less obvious analogues of “Vixens are female foxes.” The suggestion is that an armchair methodology is appropriate to their quest because it concerns truths in some sense less substantial, less world-involving than those of other disciplines: in Humean terms, relations of ideas rather than matters of fact. Our conceptual or linguistic competence, retained in the armchair, is to suffice for *a priori* knowledge of the relevant truths.

As already argued, philosophical truths are not generally truths *about* words or concepts. However, analytic truths are not supposed to be always about words or concepts, even if words or concepts are supposed to play a special role in explaining their truth. The sentence “Vixens are female foxes” is in no useful sense about the word

<sup>1</sup> To give just one example, even Jack Smart, whose work robustly engages the nature of the non-linguistic, non-conceptual world and who described metaphysics as “a search for the most plausible theory of the whole universe, as it is considered in the light of total science” (1984: 138), could also write that philosophy is “in some sense a *conceptual* inquiry, and so a science can be thought of as bordering on philosophy to the extent to which it raises within itself problems of a conceptual nature” (1987: 25), although he admits that he “cannot give a *clear* account of what I have meant when earlier in this essay I have said that some subjects are more concerned with “conceptual matters” than are others” (1987: 32).

“vixen” or any other words; it is about vixens, if anything. Its meaning is not to be confused with that of the metalinguistic sentence “‘Vixens are female foxes’ is true.” Similarly, the thought *vixens are female foxes* is not about the concept *vixen* or any other concepts; it too is about vixens, if anything. It is not to be confused with the metaconceptual thought *the thought VIXENS ARE FEMALE FOXES is true*.

How can a sentence which comes as close as “Vixens are female foxes” do to being a definition of “vixen” be about vixens rather than about the word “vixen”? Uttering it in response to the question “What does ‘vixen’ mean?” normally enables the questioner to work out the answer to the question, by pragmatic reasoning, even though the literal meaning of the sentence does not directly answer the question, just as does uttering “That is a gnu” while pointing at one in answer to the question “What does ‘gnu’ mean?.” If core philosophical truths are analytic, they may *exhibit* significant features of words or concepts without *describing* them.

Does the conception of philosophical truths as analytic or conceptual vindicate a form of the linguistic or conceptual turn without misrepresenting the subject matter of philosophy as itself linguistic or conceptual? The case study in the previous chapter gave no support to such a conjecture. Nevertheless, let us examine the matter more systematically.

Many philosophically relevant truths are clearly not conceptual truths in any useful sense. For instance, in arguing against subjective idealism, a defender of common sense metaphysics says that there was a solar system millions of years before there was sentient life. Similarly, a defender of common sense epistemology says that he knows that he has hands; that he knows that he has hands is no conceptual truth, for it is consistent with all conceptual truths that he lost them in a nasty accident. Some philosophers of time argue that not only the present exists by appeal to Special Relativity. Philosophers of mind and language dispute whether there is a language of thought; whatever the answer, it is no conceptual truth. Naturalists and anti-naturalists dispute whether there is only what there is in space and time; again, the answer is unlikely to be a conceptual truth. Moral and political philosophers and philosophers of art appeal to empirically discovered human cognitive limitations, and so on. Such philosophical arguments cannot be dismissed on general

methodological grounds. One must engage with them on their merits, in the normal way of philosophy.

Despite such examples, philosophy may be thought to have a central core of truths which are all conceptual; perhaps the rest of philosophy counts as such through its relation to the central core. Let us charitably read this restriction into the appeal to analyticity or conceptual truth in the epistemology of philosophy.

Notoriously, the idea of analyticity has been under a cloud ever since Quine argued that “a boundary between analytic and synthetic statements simply has not been drawn” (1951: 34). Nevertheless, the idea is still active in contemporary philosophy, often under the less provocative guise of “conceptual truth.” The terms “analytic” and “conceptual” will henceforth be used interchangeably.

Quine’s arguments are generally found much less compelling than they once appeared. Although he may succeed in showing that “analytic” is caught in a circle with other semantic terms, such as “synonymous,” he does not adequately motivate his jump from that point to the conclusion that the terms in the circle all lack scientific respectability, as opposed to the contrary conclusion that they all have it. Given any science, someone may insist that it define its terms, and the terms used to define them, and so on until it is driven round in a circle. By itself, that hardly demonstrates the illegitimacy of the science. Every discipline must use undefined terms somewhere or other. “Two Dogmas of Empiricism” does not explain why we should regard the undefined terms of semantics as worse off than the undefined terms of other disciplines, except by dogmatic charges of unclarity. After all, semantics is now a thriving branch of empirical linguistics. It is not to be trashed without very good reason.<sup>2</sup>

Some terms may be so unclear by ordinary working standards that no circle of definitions will render them scientifically useful. But semantic terms are not like that. By ordinary working standards, the word “synonymous” is quite clear enough to be useful. Although it is not perfectly precise – surely it has borderline cases – its degree of vagueness seems no worse than that of undefined terms in many other

<sup>2</sup> The overall criticism of Quine’s procedure goes back to Grice and Strawson (1956). Sober (2000) argues that Quine violates his own methodological naturalism in criticizing semantic notions on foundational grounds without considering their use in science.

sciences. When clarification is needed in some specific respect, it can be achieved by stipulation or otherwise, as elsewhere in science. Indeed, few contemporary philosophers feel special qualms in using the term “synonymous.” Thus any objection they have to “analytic” can hardly be based on Quine’s arguments, since his only objection to defining “analytic” in terms of “synonymous” is to the use of “synonymous” (1951: 24, 35).

The feeling remains that “analytic,” unlike “synonymous,” carries obsolescent philosophical baggage. For “analytic,” unlike “synonymous,” was once a central term in philosophical theorizing, notably in the work of logical positivists, such as Carnap, and of postwar linguistic philosophers, such as Strawson. The reason why it cannot recover that position lies not in Quine’s critique, which no longer seems compelling, but rather in Kripke’s widely accepted clarification of the differences between analyticity, apriority and necessity. Kripke did not deny that there is a boundary between the analytic and the synthetic; he merely distinguished it from other boundaries, such as the epistemological boundary between the *a priori* and the *a posteriori* and the metaphysical boundary between the necessary and the contingent (Kripke 1980: 39). He stipulated that “analytic” entails both “*a priori*” and “necessary.” Since he argued that neither of “*a priori*” and “necessary” entails the other, he was committed to denying that either of them entails “analytic” (by the transitivity of entailment).<sup>3</sup> Thus “analytic” does neither the purely epistemological work of “*a priori*” nor the purely metaphysical work of “necessary.” Its current role inevitably looks less central than the one it occupied when “*a priori*” and “necessary” were treated as pretty much

<sup>3</sup> Given Kripke’s arguments, defining “analytic” as the conjunction of “*a priori*” and “necessary” does not yield a natural notion, since a disjunction of an *a priori* contingency with an unrelated *a posteriori* necessity will then count as analytic: it is *a priori* because its first disjunct is and necessary because its second disjunct is. One does somewhat better by defining “analytic” as “*a priori* necessary,” which excludes that example, although the point of such a combination of epistemological and metaphysical elements remains to be explained. The arguments below apply to this notion too. Of course, Kripke’s main concern is the difference between the *a priori / a posteriori* and the necessary/contingent distinctions; he clarifies their differences from the analytic/synthetic distinction in passing. Nevertheless, the differentiation between the first two distinctions forces the demotion of the third from that of trying to play both the first role and the second.

interchangeable and “analytic” was taken to do the work of both. But that does not yet imply that no work remains for it to do.

If we try to sort sentences as “analytic” or “synthetic” in the manner of chicken-sexers, we can usually achieve a rough consensus. Of course borderline cases will occur, but so they do for virtually every distinction worth making: perfect precision is an unreasonable demand. The issue is what theoretical significance, if any, attaches to the rough boundary thus drawn. Even if “analytic” is defined in terms of “synonymous” and other expressions under better control than “analytic,” we should not assume without checking that it has any of the consequences sometimes associated with it. In particular, we should not assume that analytic truths are insubstantial in any further sense.

Nothing in this book challenges the legitimacy of familiar semantic terms such as “synonymous.” They will be used without apology, and they permit various senses of “analytic” to be defined. But none of them makes sense of the idea that analytic truths are less substantial than synthetic ones, or that core philosophical truths are less substantial than the truths of most other disciplines. There is something robust about “Two Dogmas of Empiricism”: insights remain even when its skepticism towards meaning is stripped away.

On some conceptions, analytic sentences are true simply in virtue of their meaning, and analytic thoughts simply in virtue of their constituent concepts. They impose no constraint on the world, not even on that part of it which consists of words and concepts. That is why it is unnecessary to get up out of one’s armchair to investigate whether such a constraint is met. Analytic truths are less substantial than synthetic ones because the latter do impose constraints on the world, which it may or may not meet. This is another way of putting the idea that analytic truths are true in virtue of meaning alone while synthetic truths are true in virtue of a combination of meaning and fact, for if analytic truths did impose constraints on the world, they would be true partly in virtue of the fact that the world met those constraints, and so not true in virtue of meaning alone. Call such conceptions of analyticity *metaphysical*. Other conceptions dispense with the idea of truth in virtue of meaning, and treat analyticity as a privileged status in respect of knowledge or justification which a sentence or thought has in virtue of the conditions for understanding its constituent words or possessing its constituent concepts. Although the privileged truths

impose constraints on the world, the task of checking that they are met is somehow less substantial than for other truths, for those who understand the relevant words or possess the relevant concepts. Call such conceptions of analyticity *epistemological*.<sup>4</sup>

This chapter examines a variety of attempts to develop a metaphysical account of analyticity. Some depend on misconceptions about meaning or truth. Others yield intelligible notions of analyticity, by watering down the traditional account to a point where it loses most of its usually supposed implications. They provide no reason to regard analytic truths as in any way insubstantial.<sup>5</sup> Even if core philosophical truths are analytic in such a sense, that does not explain how we can know or justifiably believe them.<sup>6</sup> At best it reduces the problem to the epistemology of another class of truths, such as necessary truths or logical truths. The next chapter will examine attempts to develop an epistemological account of analyticity, also with negative results. The overall upshot is that philosophical truths are analytic at most in senses too weak to be of much explanatory value or to justify conceiving contemporary philosophy in terms of a linguistic or conceptual turn.

The conclusion is not best put by calling purportedly analytic truths “substantial,” because in this context the term “substantial” is hopelessly vague. Rather, appeals in epistemology to a metaphysical conception of analyticity tend to rely on a *picture* of analytic truths as imposing no genuine constraint on the world, in order to

<sup>4</sup> See Boghossian (1997) for the distinction between metaphysical and epistemological accounts of analyticity, and Tappenden (1993: 240) for a somewhat similar distinction.

<sup>5</sup> Etchemendy (1990: 107–24) contrasts “substantive” generalizations with logical ones. The idea is widespread. It occurs in different forms in Wittgenstein’s *Tractatus Logico-Philosophicus* and in Locke’s “Of trifling propositions” (*An Essay Concerning Human Understanding*, Book IV, Chapter viii).

<sup>6</sup> Since analytic truths are standardly taken to be sentences, the term “true” will sometimes be applied to sentences, as well as to thoughts and propositions; where required, the context makes clear what kind of truth-bearer is intended. Talk of knowing or believing a sentence should be understood as elliptical for talk of having knowledge or belief which one can express with the sentence (on its standard meaning). Thus someone who knows “Grass is green” knows that grass is green and can express that knowledge by saying “Grass is green”; this is not to be confused with the meta-linguistic knowledge that the sentence “Grass is green” is true.

explain the supposed fact that knowing them poses no serious cognitive challenge. If that account could be made good, it would provide a useful sense for “insubstantial,” which would refer to the pictured property, epistemological not in its nature but in its explanatory power. Substantial truths would be the ones that lacked this property. But the account cannot be made good. The metaphysical picture cannot be filled in so as to have the required explanatory power in epistemology. Thus “substantial” and “insubstantial” are not provided with useful senses. The negation of a picture is not itself a picture. That is a problem for appeals to metaphysical analyticity, not for the present critique.

## 2

The distinction between analytic truth and synthetic truth does not distinguish different *senses* of “true”: analytic and synthetic truths are true in the very same sense of “true.” That should be obvious. Nevertheless, it is hard to reconcile with what many logical positivists, Wittgensteinians and others have said about analytic truths. For they have described them as stipulations, implicit definitions (partial or complete), disguised rules of grammar and the like. On such a conception, enunciating an analytic truth is not stating a fact but something more like fixing or recalling a notation: even if talk of truth as correspondence to the facts is metaphorical, it is a bad metaphor for analytic truth in a way in which it is not for synthetic truth. In the face of this conception, we should remind ourselves why “truth” is quite unequivocal between “analytic truth” and “synthetic truth.”

We can start by considering a standard disquotational principle for truth (where both occurrences of “P” are to be replaced by a declarative sentence):

(T) “P” is true if and only if P.

If “true” is ambiguous between analytic truth and synthetic truth, (T) must itself be disambiguated. Nevertheless, the left-to-right direction holds for both notions:

(Talr) “P” is analytically true only if P.  
(Tslr) “P” is synthetically true only if P.

Obviously, “Bachelors are unmarried” is analytically true only if bachelors are unmarried, just as “Bachelors are untidy” is synthetically true only if bachelors are untidy. The exact parallelism of (Talr) and (Tslr) already casts doubt on the supposed ambiguity. Indeed, they are jointly equivalent to a single principle about the disjunction of analytic truth and synthetic truth (“simple truth”):

(Taslr) “P” is analytically true or synthetically true only if P.

Worse, the right-to-left direction fails for both notions:

(Tarl) “P” is analytically true if P.  
(Tsrl) “P” is synthetically true if P.

For (Tarl) has a false instance when a synthetic truth is substituted for “P”; (Tsrl) has a false instance when an analytic truth is substituted for “P.” There are no natural substitutes for the right-to-left direction of (T) in the form of separate principles for analytic truth and synthetic truth. Rather, the natural substitute for the right-to-left direction disjoins the two notions:

(Tasrl) “P” is analytically true or synthetically true if P.

But (Tasrl) reinstates simple truth as the theoretically important characteristic.

One cannot avoid the problem by qualifying “true” in (T) with “analytic” for “the relevant kind of sentence” and with “synthetic” for the rest. For the sentences of the relevant kind are presumably just the analytic truths and analytic falsehoods. Thus the schemas for analytic and synthetic truth amount to these:

(Ta) If “P” is analytically true or analytically false, then “P” is analytically true if and only if P.  
(Ts) If “P” is neither analytically true nor analytically false, then “P” is synthetically true if and only if P.

But (Ta) and (Ts) follow from (Taslr), (Tasrl) and the analogue for falsity of (Taslr):<sup>7</sup>

(Faslr) “P” is analytically false or synthetically false only if not P.

Thus the information in (Ta) and (Ts) is in effect just information about the disjunction of analytic truth and synthetic truth. The attempt to treat analytic truth and synthetic truth separately just confuses the theory of “true.” The same happens for other theoretically important applications of “true.”

Consider the standard two-valued truth-table for the material conditional:

A	B	$A \rightarrow B$
T	T	T
T	F	F
F	T	T
F	F	T

If “true” is ambiguous between analytic truth and synthetic truth, what does “T” mean in that table? We might try subscripting it as  $T_{\text{analytic}}$  and  $T_{\text{synthetic}}$ , multiplying the possibilities in the first two columns accordingly and adding the appropriate subscript in the third column. “F” will require corresponding subscripts too. Since the possibilities  $T_{\text{analytic}}$ ,  $T_{\text{synthetic}}$ ,  $F_{\text{analytic}}$  and  $F_{\text{synthetic}}$  arise for both **A** and **B**, the new truth-table will have sixteen lines. Worse, consider this case:

<sup>7</sup> Proof: Assume (Taslr), (Faslr) and (Tasrl). To derive (Ta), note that it is equivalent to the conjunction of two claims: (i) if “P” is analytically true, then “P” is analytically true if and only if P; (ii) if “P” is analytically false, then “P” is analytically true if and only if P. Now (i) is logically equivalent to the claim that “P” is analytically true only if P, which follows from (Taslr). Moreover, by (Faslr) “P” is analytically false only if not P; as just seen “P” is analytically true only if P, so “P” is analytically false only if “P” is not analytically true; thus if “P” is analytically false then both sides of the biconditional in the consequent of (ii) fail, so (ii) holds. To derive (Ts), first note that “P” is synthetically true only if P by (Taslr). Conversely, if P then “P” is analytically true or synthetically true by (Tasrl); since by the antecedent of (Ts) it is not analytically true, it is synthetically true. Incidentally, by themselves (Ta) and (Ts) are weak in other ways too; in particular, they do not entail that nothing can be both analytically true and synthetically true.

A	B	$A \rightarrow B$
$T_{\text{synthetic}}$	$T_{\text{synthetic}}$	$T_{?}$

What subscript is appropriate for the third column? Suppose that Barbara is a barrister, and therefore a lawyer. Of the following four sentences, (1), (2) and (4) are synthetic while (3) is analytic (with “if” read as  $\rightarrow$ ):

- (1) Barbara is a barrister.
- (2) Barbara is a lawyer.
- (3) If Barbara is a barrister, Barbara is a lawyer.
- (4) If Barbara is a lawyer, Barbara is a barrister.

Since Barbara could easily not have been a lawyer at all, (1) and (2) are synthetic. If there are analytic truths, (3) is one of them; “barrister” simply means a lawyer with certain qualifications. Thus we cannot put “synthetic” for the missing subscript in that line of the truth-table, for that gives the wrong result when we read A as (1) and B as (2). Since Barbara could easily have been a lawyer without being a barrister, by being a solicitor, (4) is synthetic too. Thus we also cannot put “analytic” for the missing subscript, since that gives the wrong result when we read A as (2) and B as (1). Therefore the truth-table cannot be completed. Whether a material conditional is analytically true and whether it is synthetically true are not a function of whether its antecedent is analytically true, whether its antecedent is synthetically true, whether its consequent is analytically true and whether its consequent is synthetically true.

The best we can do is to put the disjunction of  $T_{\text{analytic}}$  and  $T_{\text{synthetic}}$  in the third column. But then in order to apply the truth-table iteratively, when one occurrence of  $\rightarrow$  is embedded inside another, we shall need further lines in which such disjunctions appear in the first two columns as well as the third. In effect, we have merely recovered a single sense of “true,” applicable to both analytic truths and synthetic truths, albeit awkwardly defined by a disjunction. The same conclusion can be reached by looking at combinations of other logical constants, such as conjunction and negation. What does the central work in the compositional semantics is that indiscriminate notion of truth, not the more specific notions of analytic truth and synthetic truth.

A corresponding result holds for the theory of logical consequence. Valid arguments preserve truth from premises to conclusion. What can we say if “truth” must be disambiguated between analytic truth and synthetic truth? A valid argument whose premise is a synthetic truth may have either a synthetic truth or an analytic truth as its conclusion. For example, the conjunction of a synthetic truth with an analytic truth is itself a synthetic truth, and has each conjunct as a logical consequence. For logic, the significant generalizations concern the indiscriminate disjunction of analytic truth with synthetic truth, not either disjunct separately.<sup>8</sup>

Analytic truths and synthetic truths are true in exactly the same central sense of “true.” That is compatible with their being true in very different ways, just as being a mother and being a father are two very different ways of being a parent; “parent” is not ambiguous between mothers and fathers. But truth-conditional semantics undermines even that idea. For how are (3) and (4) true in very different ways? Each is a material conditional; the antecedent and consequent of each are true in relevantly the same way as the antecedent and consequent of the other respectively. Their compositional semantic evaluation proceeds in parallel. Yet (3) is analytic, (4) synthetic. From the perspective of compositional semantics, the analytic-synthetic distinction is no distinction between different ways of being true; it is just a distinction between some truths and others.

### 3

On the metaphysical conception, analytic truths differ from synthetic ones by being true “in virtue of meaning.” The intended contrast seems to be this. A synthetic truth is true because it means what it does and things are as that meaning requires. For example, “Barbara is a barrister” is true because it means that Barbara is a barrister, and Barbara *is* a barrister. For an analytic truth, the second conjunct drops out. “Barristers are lawyers” is true simply because it means that barristers are lawyers. Nothing else is needed. But the contrast is unconvincing. For that explanation of the truth of “Barristers are lawyers” works only when we take for granted that barristers *are*

<sup>8</sup> For related arguments see Williamson (1994b: 141–2) and Tappolet (1997).

lawyers. It is no good to say “Never mind whether barristers *are* lawyers; ‘Barristers are lawyers’ is true simply because it means that barristers are lawyers.” For any true sentence *s* whatsoever, a canonical explanation of the truth of *s* takes the overall form “*s* means that P, and P.”<sup>9</sup> To use the obscure locution “in virtue of,” every true sentence is true in virtue of both its meaning and how things are. This is another way of making the point that analytic truths and synthetic truths are not true in radically different ways.<sup>10</sup>

We can ask “in virtue of” questions about non-metalinguistic matters too. In virtue of what are vixens female foxes? To use another obscure locution, what makes it the case that vixens are female foxes? An appeal to semantic or other facts about the words “vixen,” “female” and “fox” in answer to those questions would confuse use and mention. Vixens would have been female foxes no matter how we had used words. Presumably, vixens are female foxes in virtue of whatever female foxes are female foxes in virtue of; what makes it the case that vixens are female foxes is whatever makes it the case that female foxes are female foxes. Some may argue that female foxes are not female foxes in virtue of anything; nothing makes it the case that female foxes are female foxes. The suggestion may be that analytic truths require no truthmaker, unlike synthetic truths. An alternative suggestion is that analytic truths require truthmakers of a different kind from those of synthetic truths. Such suggestions are too unconstrained to be tractable for assessment. Still, two points stand out. First, they seem to conflict with general principles of

<sup>9</sup> See Boghossian (1997: 335–6). Quine says that we can say that the logical truth “Everything is self-identical” depends for its truth “on an obvious trait, viz., self-identity, of its subject matter, viz., everything.” However, he claims that it makes no difference whether we say that or say that it depends for its truth “on traits of the language (specifically on the usage of “=”), and not on traits of its subject matter” (1966: 106).

<sup>10</sup> Another problem for the supposed contrast is that it seems to equivocate on “means.” When we explain why “Barbara is a barrister” is true by saying “It means that Barbara is a barrister, and Barbara *is* a barrister,” “means” can be paraphrased as “expresses the proposition”; what proposition a sentence expresses may depend on the context in which it is uttered, if indexicals are present. By contrast, the appeal to meaning in the case of analytically true sentences is not to the proposition expressed on some particular occasion but rather to the linguistic meaning of the sentence, which is invariant across contexts, even if indexicals are present.

truthmaker theory (in the unlikely event that such a theory is needed). For instance, what makes a disjunction true is what makes one of its disjuncts true. Thus whatever makes (2) (“Barbara is a lawyer”) true also makes both (5) and (6) true:

- (5) Barbara is a lawyer or Barbara is not a lawyer.
- (6) Barbara is a lawyer or Barbara is a doctor.

But (5) is a simple logical truth, while (6) is a straightforward synthetic truth. Second, no connection has been provided between truthmaker theory and epistemology. Knowing a truth need not involve knowing its truthmaker; one can know (6) without knowing which disjunct is true (Barbara works in a building where only lawyers and doctors work). No account has been given as to why it should be easy from an armchair to know a truth with no truthmaker, or a truthmaker only of the special sort supposedly appropriate to analytic truths.

Nevertheless, at least one clear difference between paradigms of “analytic” and paradigms of “synthetic” is in the vicinity. For meaning that barristers are lawyers is sufficient for being true, whereas meaning that Barbara is a barrister is not. More generally, call a meaning *sufficient for truth* just in case necessarily, in any context any sentence with that meaning is true.<sup>11</sup> Thus the meaning of “Barristers are lawyers” is sufficient for truth; the meaning of “Barbara is a barrister” is not. One proposal is to explicate “analytic truth” as “truth whose meaning is sufficient for truth.” Call this “modal-analyticity.”<sup>12</sup> For non-skeptics about meaning and necessity, the

<sup>11</sup> To handle ambiguity, treat it as homonymy: distinct sentences with the same superficial form. The reification of meanings in the definition can be eliminated at the cost of circumlocution. Note also that the utterance of a modal-analytic truth may be false if the context shifts during the utterance: consider “If it is now exactly noon then it is now exactly noon.” Similarly, an utterance of “If John is a bachelor then John is unmarried” may express a falsehood if the wedding ceremony is completed between the utterance of the antecedent and the utterance of the consequent. Taking such complications into account would not help friends of analyticity.

<sup>12</sup> The notion of modal-analyticity is similar to the notion of deep necessity in Evans (1979), where the truth of the sentence does not depend on any contingent feature of reality.

notion of modal-analyticity is quite intelligible. But what are its consequences?

Consider any non-indexical sentence  $s$  that expresses a necessarily true proposition. Necessarily, in any context, any sentence with the actual meaning of  $s$  expresses that necessary truth and is therefore true. Thus  $s$  is a modal-analytic truth, because its meaning is sufficient for truth. In that sense, it is true in virtue of meaning. But how little has been achieved in so classifying it! Nothing has been done to rule out the hypothesis that it expresses a profound metaphysical necessity about the nature of the world, knowable if at all only through arduous *a posteriori* investigation, for instance. No reason has been provided to regard  $s$  as “merely verbal” or “insubstantial” in a pretheoretic sense, unless one already had independent reason to regard all necessities as merely verbal or insubstantial. Similarly, mathematical truths count as modal-analytic; their so counting is by itself no reason to regard them as merely verbal or insubstantial. Indeed, for all that has been said, even “Water contains H<sub>2</sub>O” is modal-analytic, given that “water” has a different meaning as used on Twin Earth to refer to XYZ, a different substance with the same superficial appearance.

To make the point vivid, call a meaning *temporally sufficient for truth* just in case at all times, in any context any sentence with that meaning is true. Read the quantifiers “at all times” and “in any context” non-modally, so they do not range outside the actual world. Thus any sentence which expresses, in a time-independent way, an eternally true proposition, however contingent, has a meaning temporally sufficient for truth. For example, the meaning of “No hotel ever has a billion rooms” is presumably temporally sufficient for truth. We can call the sentence “temporal-analytic” if we like, but that in no way implies that it is somehow insubstantial, because there is no background connection between eternity and some sort of insubstantiality. Similarly, calling a sentence “analytic” in the sense of modal-analyticity does not imply that it is somehow insubstantial, in the absence of a background connection between necessity and some sort of insubstantiality. Yet the account of analyticity was what was supposed to substantiate the claim of insubstantiality. If we already had a background connection between necessity and insubstantiality, there would be little to gain from invoking modal-analyticity in order to argue that core philosophical truths are insubstantial, since

we could do it more simply just by arguing that true philosophical sentences in the core express necessarily true propositions.

Admittedly, not all modal-analytic true sentences express necessarily true propositions. Examples of the contingent *a priori* such as “It is raining if and only if it is actually raining” are modal-analytic, since the truth of “It is raining” as uttered in a given context is necessarily equivalent to the truth of “It is actually raining” as uttered in that context, because “actually” refers rigidly to the world of the context, but the biconditional does not express a necessary truth, since the weather could have been relevantly different, in which case it would have been not raining if and only if it is actually raining. Thus modal-analyticity violates Kripke’s constraint that analyticity implies necessity; in this respect it may diverge from the traditional conception. Conversely, not all sentences that express necessarily true propositions are modal-analytic: consider examples of the necessary *a posteriori* such as “I am not Tony Blair.” Nevertheless, such examples seem marginal to the envisaged conception of core philosophical truths, most of which will both express necessarily true propositions and be modal-analytic.

A core of philosophical truths may indeed be modal-analytic. Some philosophers seek to articulate necessary truths without essential reliance on indexicals; if they succeed, the sentences they produce are modal-analytic. Even if contextualists are right, and key philosophical terms such as “know” shift their reference across contexts, the relevant sentences may still both express necessarily true propositions and be modal-analytic: consider “Whatever is known to be the case is the case.” The answers to philosophical questions of the forms “Is it possible that P?” and “Is it necessary that P?” will themselves express necessary truths, given the principle of the widely accepted modal logic S5 that the possible is non-contingently possible and the necessary non-contingently necessary; if the answers can be phrased in non-indexical terms, they will then be modal-analytic. But outside the envisaged core many philosophically relevant truths will not be modal-analytic, as the examples near the start of the chapter show.

Unfortunately, even for modal-analytic philosophical truths, classifying them as modal-analytic does not unlock their epistemology, any more than classifying a truth as necessary explains how we can know it. Of course, if a sentence is modal-analytic, then one is safe from error in uttering it with its given meaning. In that sense, one’s utterance is reliable. But such reliability falls well short of what

knowledge requires, since otherwise any true mathematical assertion would count as an expression of knowledge, no matter how fallacious the “proof” on which it was based. “Vixens are female foxes” is utterly misleading as a paradigm for the epistemology of modal-analytic truths in general. To say that *s* is a modal-analytic truth whose constituent words and grammar we understand does very little way to explain how we can know or justifiably believe *s*.<sup>13</sup> In particular, it does not imply that the mere linguistic understanding of *s*, which every competent speaker possesses, provides any insight into the truth of *s*, or constitutes more than the minimal starting-point for inquiry it does for ordinary synthetic truths.

## 4

Issues related to those just raised for modal-analyticity arise for what is sometimes called “Frege-analyticity.”<sup>14</sup> A sentence is Frege-analytic just in case it is synonymous with a logical truth. For example, “All furze is furze” is a logical truth, roughly speaking because everything of the form “All *F* is *F*” is true. “All furze is gorse” is not a logical truth, because not everything of the form “All *F* is *G*” is true (“All fungus is grease” is false). However, “All furze is gorse” is Frege-analytic, because it is synonymous with the logical truth “All furze is furze,” since “furze” is synonymous with “gorse.” In “Two Dogmas,” Quine admits the notion of logical truth, and therefore allows that if “synonymous” were legitimate, so would be “analytic” in the sense of Frege-analyticity. By present standards, the notion of Frege-analyticity is quite intelligible. But what are its consequences?

Trivially, every logical truth is Frege-analytic, because it is synonymous with itself. Clearly, this alone does nothing to show that logical truths are somehow insubstantial in any metaphysical, epistemologically explanatory sense (see the end of Section 1). For instance, it is compatible with the hypothesis that there are truths of second-order logic which characterize the necessary structure of reality in profound

<sup>13</sup> See n. 6 for this terminology.

<sup>14</sup> The term “Frege-analytic” is from Boghossian (1997), with reference to §3 of Frege (1950) (as Boghossian suggests, the interpretation of the passage is not entirely clear). He classifies the notion of Frege-analyticity as neither epistemological nor metaphysical but semantic (1997: 363); for convenience, it is treated here under the heading of metaphysical notions of analyticity.

ways and can never be known by any mind. A *fortiori*, nothing has been done to show that Frege-analytic truths are insubstantial.<sup>15</sup>

To make the point vivid, call a sentence “Einstein-analytic” just in case it is synonymous with a truth once uttered by Einstein. Trivially, every truth once uttered by Einstein is Einstein-analytic. That does nothing to show that truths once uttered by Einstein are in any sense insubstantial; a *fortiori*, nothing has been done to show that Einstein-analytic truths are somehow insubstantial. Of course, if we had independent reason to regard all logical truths as somehow insubstantial, that would presumably give us reason to regard all Frege-analytic truths as insubstantial in some related way, but the mere definition of “Frege-analytic” provides no such reason. Quine devoted some of his most powerful early work to arguing that logical truths are not analytic in a less trivial sense (Quine 1936).

To explain why “All furze is furze” is a logical truth while “All furze is gorse” is not, use was made of Tarski’s standard model-theoretic account of logical consequence as truth-preservation under all interpretations which preserve logical form, and in particular of logical truth as truth under all such interpretations (Tarski 1983b). It lends no support to any conception of logical truths as somehow insubstantial. The truth of a sentence under all interpretations which preserve its logical form in no way make its truth under its intended interpretation insubstantial.<sup>16</sup> To use a style of argument from Section 2, consider this simple logical truth (with “if” read as the material conditional):

(7) If Barbara is a barrister, Barbara is a barrister

Its compositional semantic evaluation proceeds in parallel to that for the non-logical analytic truth (3) and the synthetic truth (4); each is true because it is a material conditional with a true antecedent and a true consequent. All three are true in the same way. From the perspective of compositional semantics, logical truths are true in the same way as other truths.

In one good sense, sentences of the form “P if and only if actually P” are logical truths, and therefore Frege-analytic, because true in

<sup>15</sup> Quine (1966: 111) notes that so-called truth by definitions (“Every vixen is a female fox”) depends on prior logical truths (“Every female fox is a female fox”).

<sup>16</sup> Note that the epistemological issue is not how we can know that *s* is a logical truth; it is how, given that *s* is a logical truth, we can know the simple truth of *s*.

every model (Davies and Humberstone 1980, Kaplan 1989). Nevertheless, they can express contingent truths on the same reading; it is not necessary for me to be my actual height. Although we could add a modal qualification to the definition of logical truth in order to exclude such examples, by requiring logical truths to be true at every world in every model, this mixing together of the modal dimension with the world dimension is bad taxonomy; perspicuous basic notions keep such different dimensions separate. Thus Frege-analyticity, like modal-analyticity, violates Kripke's constraint that analyticity implies necessity. In this respect Frege-analyticity too may diverge from the traditional conception.

The mathematical rigor, elegance, and fertility of model-theoretic definitions of logical consequence depend on their freedom from modal and epistemological accretions. As a result, such definitions provide no automatic guarantee that logical truths express necessary or *a priori* propositions. This is no criticism. As a theoretical discipline, logic only recently attained maturity. Tarski's model-theoretic notion of logical consequence has turned out to be a key theoretical notion. To reject it on the basis of preconceived extraneous constraints would subvert the autonomy of logic as a discipline. Pretheoretic conceptions of logical consequence are in any case too confused to provide much guidance on subtle issues.<sup>17</sup> Still, those who do have a non-standard account of logical truth can feed it into the definition of "Frege-analytic" if they like.

"All furze is furze," unlike many logical truths, is obvious. That does not justify the idea that it imposes *no* constraint on the world, rather than one which, by logic, we easily know to be met (Wittgenstein, *Tractatus Logico-Philosophicus*, 4.461–4.4661 and 6.1–613). What case does the constraint exclude? That not all furze is furze, of course. To complain that "Not all furze is furze" does not express a genuine case is to argue in a circle. For it is to assume that a genuine constraint must exclude some logically consistent case. Since substantiality was being understood to consist in imposing a genuine constraint, that is tantamount to assuming that no logical truth is substantial, the very point at issue. Concentration on obvious logical truths obscures this circularity.

<sup>17</sup> For more discussion and further references to the controversy over the nature of logical consequence see Williamson (2000b).

We may hope, given an epistemology for logical truths, to extend it to an epistemology for Frege-analytic truths. That task will not be trivial, for cognitive differences may arise between synonymous expressions, even for those who understand them. For example, Kripke (1979) has argued persuasively that a competent speaker of English can understand the synonymous expressions “furze” and “gorse” in the normal way without being in a position to know that they refer to the same thing. Such a speaker will assent to the logical truth “All furze is furze” while refusing assent to the Frege-analytic truth “All furze is gorse.” Similarly, on standard theories of direct reference, coreferential proper names such as “Hesperus” and “Phosphorus” are synonymous, so an astronomically ignorant competent speaker may assent to the logical truth “If Hesperus is bright then Hesperus is bright” while refusing assent to the Frege-analytic truth “If Hesperus is bright then Phosphorus is bright.”

The epistemological consequences of such examples are contested. According to some direct reference theorists, the proposition that if Hesperus is bright then Phosphorus is bright *is* the proposition that if Hesperus is bright then Hesperus is bright, so whoever knows that if Hesperus is bright then Hesperus is bright *ipso facto* knows that if Hesperus is bright then Phosphorus is bright.<sup>18</sup> However, even granted that view of propositional attitude ascriptions, that speaker is in no position to know that if Hesperus is bright then Phosphorus is bright under the guise of the sentence “If Hesperus is bright then Phosphorus is bright,” but only under the guise of the sentence “If Hesperus is bright then Hesperus is bright.” In a sense the speaker cannot express their knowledge by using the merely Frege-analytic sentence, even though it expresses the content of that knowledge: if they do use the sentence, their utterance will not be causally connected to their knowledge state in the right way. In elliptical terms, the speaker knows “If Hesperus is bright then Hesperus is bright” without being in a position to know “If Hesperus is bright then Phosphorus is bright”; they know the logically true sentence without being in a position to know the merely Frege-analytically true sentence.

If propositions are individuated in that coarse-grained direct reference way, what matters for progress in philosophy is less which propositions we know than which sentential guises we know them under. Suppose, just for the sake of argument, that some form of

<sup>18</sup> See Salmon (1986), especially 133–5.

physicalism is true, and pain is in fact identical with  $\pi$ , where “ $\pi$ ” is a name whose reference is fixed by a neuroscientific description. According to a hard-line direct reference theory, “pain” and “ $\pi$ ” are synonymous. The hypothesis “Pain is  $\pi$ ” becomes a focus of philosophical controversy. On some direct reference theories, everyone knew all along that pain is  $\pi$ , because they knew all along that pain is pain and the proposition that pain is  $\pi$  just is the proposition that pain is pain. If that view is correct, it just shows that such attitude ascriptions constitute the wrong level of description for understanding philosophical activity. What matters is that although everyone knew the proposition under the guise of the logical truth “Pain is pain,” they did not know or even believe it under the guise of the merely Frege-analytic truth “Pain is  $\pi$ .” In elliptical terms, they knew “Pain is pain” but not “Pain is  $\pi$ .” Perhaps such physicalist theories are false, but we can hardly expect philosophy to be a discipline in which there are no informative identities; the moral of the example stands. The need for such finer-grained descriptions of propositional attitudes is even more urgent if propositions as the objects of knowledge and belief are identified with sets of possible worlds, for then all necessary truths are identical with the set of all possible worlds: anyone who knows one necessary truth knows them all (Lewis 1996, Stalnaker 1999: 241–73). Thus a coarse-grained account of attitude ascriptions does not trivialize the problem of extending an epistemology for logical truths to an epistemology for Frege-analytic truths.

Opponents of direct reference theories usually hope to make synonymy a more cognitively accessible relation for competent speakers. However, the prospects for making it perfectly accessible are very dubious. Pairs such as “furze” and “gorse” are pre-theoretically plausible cases of synonymous expressions that speakers can understand in the ordinary way without being in a position to know them to be synonymous.<sup>19</sup> The extension of an epistemology for logical truths to an epistemology for Frege-analytic truths will probably have to allow for significant cognitive obstacles that cannot be overcome simply by speakers’ ordinary linguistic competence.

<sup>19</sup> See Kripke (1979). This contradicts Dummett’s claim that “It is an undeniable feature of the notion of meaning – obscure as that notion is – that meaning is *transparent* in the sense that, if someone attaches a meaning to each of two words, he must know whether these meanings are the same (1978: 131). For more general theoretical considerations against such claims see Williamson (2000a: 94–107). See also Horwich (1998: 100–1).

Even for sentential guises, identity and distinctness are not guaranteed to be transparent to speakers: someone may be confused as to whether “Paderewski,” the name of the politician, is the same name as “Paderewski,” the name of the pianist (Kripke 1979). A single speaker at a single time may associate different mental files with the same word of a natural language, or the same mental file with different words of the language. Speakers may also be confused as to whether they are calling on two mental files or one. What needs to be found is not the mythical level of description at which perfect transparency to the subject is guaranteed but rather a perspicuous level of description at which the relevant cognitive phenomena are individuated in a way that is neither so coarse-grained that the most relevant distinctions cannot be drawn nor so fine-grained that they are drowned out by a crowd of irrelevant ones. Since philosophical debates involve many interacting individuals, sentential guises usually provide an appropriate level of description.

We also need an epistemology for logical truths in the first place. To that, the notion of Frege-analyticity contributes nothing. In particular, that a sentence is Frege-analytic does not imply that mere linguistic competence provides any insight into its truth, or constitutes more than the minimal starting-point for inquiry it does for ordinary synthetic truths.

How many philosophical truths are Frege-analytic? As a simple example, take the true sentence “Persons are not events” (if you think that persons are events, take “Persons are events” instead). It is not itself a logical truth, on any standard conception of logic. In particular, “person” and “event” seem not to be logical constants, and the logical form “Ps are not Es” has false instances such as “Parisians are not Europeans.” What logical truth could “Persons are not events” be synonymous with? “Persons who are not events are not events” is a logical truth, but not synonymous with the original. Granted, “persons” and “persons who are not events” have the same intension (function from circumstances of evaluation to extension) in every context of utterance.<sup>20</sup> Still, they are not literally synonymous, for whatever the semantic structure of “persons,” it is finite, and

<sup>20</sup> The contexts of utterance and circumstances of evaluation here are not restricted to the actual world. If the content of an expression has a structure which reflects the grammatical structure of the expression, then sameness of intension does not imply sameness of content, and sameness of intension in every context does not entail

therefore a proper part of the semantic structure of “persons who are not events”; thus the two expressions differ in semantic structure. One can try to construct non-circular analyses of “person” and “event” or both whose substitution into the sentence would yield a logical truth: “To be a person is to be a *QRS*.” However, “person” and “*QRS*” are unlikely to be literally synonymous. Almost certainly, someone will produce a purported counterexample to the analysis: “Such-and-such would be a person but not a *QRS*” or “So-and-so would be a *QRS* but not a person.” Direct reference theorists will tend to expect just such counterexamples to the claim that the apparently simple term “person” and the complex description “*QRS*” have the same intension; direct reference theories partly originate from Kripke and Putnam’s counterexamples to a host of similar descriptivist claims. Opponents of direct reference may be less pessimistic about the prospects for a complex description with the same intension as “person.” However, on their finer-grained views of meaning, on which synonymy is as transparent as possible to competent speakers, a purported counterexample need not be correct to defeat the claim of synonymy: what counts is that its proponent is neither linguistically incompetent nor fundamentally irrational. Contemporary proponents of a descriptivist view of meaning as a rival to direct reference theory usually envisage a loose semantic connection with a cluster of descriptions rather than strict synonymy with a single description. Whichever side of the debate one takes, there are good grounds for skepticism about the supposed synonymy of “person” and “*QRS*.” The best bet is that “Persons are not events” is not Frege-analytic. The point does not depend on peculiarities of the example; it could be made just as well for most other philosophical claims.<sup>21</sup> In contemporary philosophy, few who propose complex analyses claim synonymy for them.<sup>22</sup>

One might react by loosening the relation of synonymy to some equivalence relation that would have a better chance of holding

---

sameness of character, that is, sameness of content in every context. See Kaplan (1989) for relevant background.

<sup>21</sup> Boghossian argues that many *a priori* truths are not Frege-analytic (1997: 338–9).

<sup>22</sup> This point is related to the paradox of analysis: how can a conceptual analysis be both correct and informative? The paradox goes back to Langford (1942).

between the *analysandum* and the *analysans* in philosophically significant analyses. Call the looser equivalence relation “metaphysical equivalence.” A wider class of philosophical truths might be transformable into logical truths by the substitution of metaphysically equivalent terms. Call the truths in the wider class “quasi-Frege-analytic.” The poor track record of philosophical analysis does not suggest that the class of quasi-Frege-analytic truths will be very much wider than the class of Frege-analytic truths.<sup>23</sup> In any case, the looser metaphysical equivalence is, the more problematic it will be to extend an epistemology for logical truths to an epistemology for quasi-Frege-analytic truths. The aim of the loosening is to permit some distance between the meaning of the *analysandum* and the meaning of the *analysans*; that will tend to make even the coextensiveness of the *analysandum* and *analysans* less cognitively accessible. There will be a corresponding tendency to make the material equivalence of the original quasi-Frege-analytic truth to the logical truth less cognitively accessible too.

For instance, one might define “metaphysical equivalence” as sameness of intension in every context. The question is then how the sameness of intension in every context of the substituted terms could enable one to advance from knowing or justifiably believing the logical truth to knowing or justifiably believing the merely quasi-Frege-analytic truth. No guarantee has been provided that we can know or justifiably believe the universally quantified biconditional of the substituted terms. By hypothesis, that biconditional will in fact express a necessary truth in every context; the problem merely shifts to how such truths can be known, just as in the case of modal-analyticity. If that problem were already solved, there would be little to gain from appealing to quasi-Frege-analyticity in order to explain how core philosophical truths can be known.

Even if many philosophical truths are quasi-Frege-analytic, it does not follow that we can gain cognitive access to them simply on the basis of our logical and linguistic competence.

Yet another proposal is to consider as (metaphysically) analytic just the logical consequences of true (or good) semantic theories. It is presumably in the spirit of this proposal to interpret semantic theories not as stating straightforwardly contingent, *a posteriori* facts about how people use words but as somehow articulating the essential structure of semantically individuated languages; in this sense, the word “green”

<sup>23</sup> See Fodor (1998: 69–87) and Williamson (2000a: 31–3) for further discussion.

could not have meant anything but *green* in English. Even so, the definition does nothing to trace any special cognitive access that speakers have to semantic facts about their own language to any special metaphysical status enjoyed by those facts. It also counts every logical truth as analytic, since a logical truth is a logical consequence of anything, without illuminating any special cognitive access we may have to logical truths. Of course, *if* someone knows the relevant semantic truths about their own language and is logically proficient, then they are also in a position to know the analytic truths as so defined. But, on this definition, we do nothing to explain how the semantics and logic are known in the first place by saying that they are analytic. As in previous cases, the account of analyticity merely shifts the burden from explaining knowledge of analytic truths to explaining knowledge of some base class of necessary or logical or semantic or other truths. Once the analyticity card has been played to effect this shift of the explanatory burden, it cannot be played again to explain knowledge of the base truths, by saying that they are analytic, for they count as analytic simply because they belong to the relevant base class, and the question remains how we know truths in the base class.

## 5

Unless one is a skeptic about meaning or modality, one can define several notions of analyticity in semantic and modal terms, but none of them provides any reason to regard the truths to which it applies as somehow insubstantial, or as posing no significant cognitive challenge. That upshot may seem puzzling. Surely we sometimes make a sentence true by stipulative definition. For example, I might introduce the term “zzz” (pronounced as a buzz) by saying “A zzz is a short sleep” and thereby make “A zzz is a short sleep” true. What prevents us from using such cases as paradigms to fix a semantic notion of analyticity on which analytic truths are insubstantial?

We can see the problems for the proposal more clearly by distinguishing the semantic from the metasemantic. Semantics facts are facts of the kind we attempt to systematize in giving a systematic compositional semantic theory for a language, facts as to what its expressions mean. Metasemantic facts are the nonsemantic facts on which the semantic facts supervene. The distinction is rough but clear enough to be workable. Thus the fact that “horse” applies to horses

is semantic, not metasemantic; the fact that utterances of “horse” are often caused by horses is metasemantic, not semantic.<sup>24</sup> Similarly, the fact that “zzz” means a short sleep is semantic, while the fact that it was introduced by someone saying “A zzz is a short sleep” is metasemantic. The semantic theory takes no notice of the act of stipulation, only of its outcome – that a given expression has a given meaning. The act of stipulation makes the sentence true by making it have a meaning on which it is, in the quite ordinary way, true. My saying “A zzz is a short sleep” did not make a zzz be a short sleep, because that would be to make a short sleep be a short sleep, and my saying “A zzz is a short sleep” certainly did not make a short sleep be a short sleep. In particular, since there were many short sleeps before I was born, there were many zzzes before I was born, independently of my later actions. At best, my saying “A zzz is a short sleep” made “zzz” mean a short sleep, and therefore “A zzz is a short sleep” mean that a short sleep is a short sleep. This is simply the standard semantic contribution of meaning to truth, just as for synthetic truths. The peculiarity of the case is all at the metasemantic level; the use of stipulative definitions as paradigms does not yield a *semantic* notion of analyticity. Making “zzz” mean a short sleep helps make “A zzz is a short sleep” true only because a short sleep is a short sleep. “A short sleep is a short sleep” is a logical truth, but we have still been given no reason to regard logical truths as somehow insubstantial. The use of stipulative definitions as paradigms of analyticity does not justify the idea that analytic truths are in any way insubstantial.

My stipulation may smooth my path from knowing the logical truth “A short sleep is a short sleep” to knowing the Frege-analytic truth “A zzz is a short sleep,” but of course that does not explain how I know “A short sleep is a short sleep” in the first place.

The metaphysics and semantics of analytic truths are no substitute for their epistemology. If their epistemology is as distinctive as is often supposed, that is not the outcome of a corresponding distinctiveness in their metaphysics or semantics. It can only be captured by confronting their epistemology directly. We therefore turn to epistemological accounts of analyticity.

<sup>24</sup> For helpful discussion see the essays in Part IV of Stalnaker (2003). He sometimes use the terminology of “descriptive semantics” and “foundational semantics” rather than “semantics” and “metasemantics” respectively.

# Epistemological Conceptions of Analyticity

---

As observed in the previous chapter, metaphysical conceptions of analyticity do not themselves imply that linguistic or conceptual competence constrains one's attitudes to analytic sentences or thoughts. If our interest is in such constraints, we had best consider them directly. We can then assess what role, if any, they play in explaining the armchair methodology of philosophy.

If someone is unwilling to assent to the sentence “Every vixen is a female fox,” the obvious hypothesis is that they do not understand it, perhaps because they do not understand the word “vixen.” The central idea behind epistemological conceptions of analyticity is that, in such cases, failure to assent is not merely *good evidence* of failure to understand; it is *constitutive* of such failure. Of course, it is not by itself constitutive of failure to understand the word “vixen”, since someone who understands that word may nevertheless not assent to the sentence, for example because they do not understand the word “fox”; a monolingual speaker of another language may understand “vixen” through the testimony of a bilingual without understanding any other word of English. Rather, failure to assent to the sentence can by itself only be constitutive of failure to understand the whole sentence. An unqualified link from understanding to assent is this:

(UAl) Necessarily, whoever understands the sentence “Every vixen is a female fox” assents to it.

One proposal is to generalize UAl to define an epistemological notion of analyticity: a sentence  $s$  is analytic just in case, necessarily, whoever

understands *s* assents to *s*. We could go further, by articulating an explicitly constitutive and not merely modal connection, but for present purposes the question is whether even this proposed necessary connection holds.

Three obvious glosses on UAl must be taken as read throughout. First, it concerns “Every vixen is a female fox” with its current meaning, for of course if the phonetically individuated sentence had meant something different, someone might easily have understood it and refused to assent. Second, assent is dispositional, for of course we are not actively assenting to any sentence whenever we understand it. Third, assent is a mental attitude, not a merely verbal one, for someone might easily understand “Every vixen is a female fox” while refusing to give it overt assent, for example because overt assent to a triviality looks uncool. We could speak of belief rather than assent, but the latter term sounds more natural in relation to inference rules, to which the notion of analyticity will be generalized.

A corresponding notion of analyticity can be defined for thoughts: a thought *t* is analytic just in case necessarily, whoever grasps *t* assents to *t*. If the thought *every vixen is a female fox* is analytic in this sense, then:

(UAt) Necessarily, whoever grasps the thought *every vixen is a female fox* assents to it.

On the simplest view, thinking a thought with any attitude towards it suffices for grasping it. Friends of principles like UAt should beware of straying too far from that simple view, by claiming that “full grasp” of a thought requires much more than the ability to think it (Peacocke 1992: 29–33, Bealer 1998: 221–2). For such a defence of UAt risks trivializing it, by in effect writing the consequent into the antecedent by hand. At any rate, grasp of a thought should be a matter of normal conceptual competence, just as understanding of a sentence is a matter of normal linguistic competence. We shall return to these issues below.

Call UAl and UAt “understanding-assent links” for language and thought respectively. The picture is that grasping a thought consists of grasping its constituent concepts and the way in which they have been put together just as understanding a sentence consists of understanding its constituent expressions and syntax.

Assent is no metalinguistic or metaconceptual attitude: normally, in actively assenting to “Grass is green,” one is saying or thinking that grass is green, not that the sentence “Grass is green” or the thought *grass is green* is true. However, thinking *grass is green* cannot be uncontroversially equated with thinking that grass is green. For thinking that grass is green presumably has as its object the proposition that grass is green. On a Russellian view, that proposition is made up of grass and greenness themselves, not of the concepts *grass* and *green*. Thus the thought *grass is green*, which is composed of concepts, must be distinguished from the proposition that grass is green. The thought is something like a mental vehicle for the proposition. Moreover, the same proposition can have different vehicles. For example, on this Russellian view, the proposition that *Hesperus*, if it exists, appears in the evening is the proposition that *Phosphorus*, if it exists, appears in the evening. The friend of conceptual connections is still likely to distinguish the concept *Hesperus* from the concept *Phosphorus*, and the thought *Hesperus, if it exists, appears in the evening* from the thought *Phosphorus, if it exists, appears in the evening*, on the grounds that the former embodies a conceptual connection while the latter does not. Thus understanding-assent links for thought must be articulated in terms of thoughts rather than propositions, in case there is a difference (for Fregeans, the proposition is the thought). Assenting to the thought *grass is green* is something like judging that grass is green under the guise of that thought. Similarly, assenting to the sentence “Grass is green,” for someone who understands it, is something like believing that grass is green under the guise of that sentence. More generally, in a context in which the sentence *s* expresses the proposition *p*, assenting to *s*, for someone who understands it, is something like believing *p* under the guise of *s*. For you, assenting to “I am hungry” is something like believing that you are hungry under the guise of the sentence “I am hungry,” since in your context that sentence expresses the proposition that you are hungry, not the proposition that I am hungry. Similarly, in a context in which the thought *t* expresses the proposition *p*, assenting to *t* is something like believing *p* under the guise of *t*.

The notion of an understanding-assent link can be generalized from individual sentences or thoughts to arguments at the level of language or thought. For example, if someone is unwilling to assent to the inference from “This is red and round” to “This is red,” the

obvious hypothesis is that they do not understand one of the sentences, most probably because they do not understand the word “and.” For epistemological conceptions of analyticity, failure to assent in such cases is again not merely good evidence of failure to understand but constitutive of such failure. Gerhard Gentzen introduced the idea that some rules of his natural deduction systems of logic have definitional status. Following him, a tradition which includes Dag Prawitz, Michael Dummett, Per Martin-Löf, Christopher Peacocke, Robert Brandom, Paul Boghossian and many others has developed in various ways the conception of acceptance of such inference rules as playing a constitutive role in understanding the logical constants, and therefore in understanding the sentences in which they occur. For many of these thinkers, this is one step towards a quite general “inferentialist” account of meaning and understanding for expressions in terms of their conceptual roles.<sup>1</sup>

Understanding-assent links, or something like them, are also commonly thought to play a leading role in the understanding of theoretical terms in science: if you don’t assent to some core sentences of electron theory, in which the word “electron” occurs, you don’t understand the word, and therefore don’t understand those sentences.

A natural project is therefore to try to explain the armchair methodology of philosophy as based on something like understanding-assent links: our sheer linguistic and conceptual competence mandates assent to some sentences or thoughts and inferences, which form the starting-point for philosophical inquiry. This chapter assesses the prospects for such a project.

The envisaged method cannot accurately be characterized as “reflection on our own concepts.” For that description specifies the method only as “reflection,” which applies to virtually all forms of philosophy. Moreover, it specifies the subject matter as “our own concepts,” whereas the envisaged method involves reflection *with* our own concepts, and is therefore reflection *on* whatever our concepts

<sup>1</sup> The case of deductive logic is a useful reminder that many short, trivial steps of no apparent philosophical significance can be chained together into a long, non-trivial argument of obvious philosophical significance. The short steps were not really philosophically insignificant after all: no apologies for concentrating on them here.

happen to refer to – in most cases, not concepts. The idea is rather to exploit whatever epistemic assets we have simply in virtue of our linguistic and conceptual competence. Suppose that a philosopher arrives at a theory about understanding, reference, and concepts by employing a battery of general armchair techniques that rely on far more than mere linguistic and conceptual competence. Say, for definiteness, that the theory gives a crude “best fit” account of reference, and entails that justice is whatever best fits our beliefs about justice. Pretend that the theory is true. Even so, it does not follow that “Justice is whatever best fits our beliefs about justice” is epistemologically analytic. For it was not reached on the basis just of linguistic and conceptual competence. Similarly, a definition of “conceptual truth” as “truth of the theory of concepts” is unhelpful for present purposes, since it merely raises the question how the truths of the theory of concepts are known (“metaconceptual truth” would be less misleading terminology).

In what follows, we will consider more rigorously what is epistemically available simply on the basis of linguistic and conceptual competence. To a first approximation, the answer is: nothing.

## 2

We start with a provisional sketch of some obstacles to extracting epistemological consequences from understanding-assent links and of some attempts to overcome them. Then we turn in Section 3 to the main argument: that understanding-assent links simply do not hold.

Our concern is knowledge or justification, not just belief or assent. On the most optimistic view, understanding-assent links generate understanding-knowledge links like these:

- (UKl) Necessarily, whoever understands the sentence “Every vixen is a female fox” knows “Every vixen is a female fox.”
- (UKt) Necessarily, whoever grasps the thought *Every vixen is a female fox* knows *Every vixen is a female fox*.

Here, knowing “Every vixen is a female fox” amounts to knowing that every vixen is a female fox under the guise of the sentence “Every

vixen is a female fox,” and knowing *every vixen is a female fox* amounts to knowing that every vixen is a female fox under the guise of the thought *every vixen is a female fox*. Since knowing something entails assenting to it (we may assume), UKl and UKt entail UAl and UAt respectively. But since assenting to something does not entail knowing it, how are understanding-knowledge links to be extracted from understanding-assent links? UAl and UAt do not entail UKl and UKt in any obvious way.

An even more elementary problem arises. Knowledge is factive. Thus understanding-knowledge links entail corresponding understanding-truth links:

- (UTl) Necessarily, someone understands the sentence “Every vixen is a female fox” only if it is true.
- (UTt) Necessarily, someone grasps the thought *Every vixen is a female fox* only if it is true.

Thus if understanding-assent links somehow imply the corresponding understanding-knowledge links, *a fortiori* they also imply the understanding-truth links. Perhaps UTl and UTt hold because the sentence “Every vixen is a female fox” and the thought *every vixen is a female fox* are necessarily true. But in other cases the question of truth becomes more urgent.

Consider theoretical terms from discredited theories. If an understanding-assent link holds for “phlogiston,” and understanding “phlogiston” necessitates assent to a core of phlogiston theory, how could it follow that someone understands sentences of phlogiston theory only if a core of it is true? Didn’t proponents of phlogiston theory understand their own theory, despite its untruth? The example is not completely straightforward, for at least two reasons. First, it requires the untruth of the core of phlogiston theory in the understanding-assent links, not just of phlogiston theory as a whole. Some will treat a universal generalization of the form “All phlogiston is . . .” as vacuously true if phlogiston does not exist. Second, if there is nothing for “phlogiston” to refer to, one might alternatively treat sentences in which it occurs as failing to express propositions, in which case it is unclear that genuine understanding of phlogiston theory is possible. For the sake of the example, however, we may suppose that a core claim of phlogiston theory is of the form “Phlo-

giston plays role R,” that a necessary condition of understanding the term “phlogiston” is assenting to that claim, and that the claim is untrue, because nothing plays role R. Suppositions of this kind will be questioned later.

We are sometimes advised to drop various ordinary terms, on the grounds that obsolete and false folk theories are built into them. Those who offer such advice may be assuming that understanding-truth links fail for some critical sentences of the folk theory in which those terms occur while the corresponding understanding-assent links hold (if so, they presumably do not count themselves as fully understanding the folk theory). For if we can understand the critical sentences of the folk theory without assenting to them, in what sense is the theory built into the key terms? For example, we could use them to assert the negations of central principles of the theory.<sup>2</sup>

Some understanding-assent links might even be to logically inconsistent sentences or thoughts. For example, the ordinary notion of truth is sometimes held to be incoherent, on the grounds that a necessary condition for understanding “true,” and so for understanding sentences in which it occurs, is assent to a disquotational principle for “true” which the Liar paradox shows to be inconsistent. Tarski’s description of natural languages as “inconsistent” in virtue of the paradox (1983a: 164–5) may involve such a view, for if we can understand “true” in English without assenting to the troublesome instances of the disquotational principle, what prevents us from using English consistently?<sup>3</sup> Similarly, Prior’s connective “tonk” has mismatched introduction and elimination rules; the introduction rule licenses the inference from “P” to “P tonk Q,” while the elimination rule licenses the inference from “P tonk Q” to “Q” (Prior 1960). By putting these rules together, one can derive any conclusion “Q” from any premise “P.” If assent to instances of those rules is necessary for understanding them, because necessary for understanding “tonk,” it hardly follows that the rules are truth-preserving (in the context of someone who understands “tonk”); they are so only if either every

<sup>2</sup> In effect, Horwich (1998: 131–53) allows understanding-belief links for which the understanding-truth links fail.

<sup>3</sup> See Eklund (2002) for a defense of the idea of inconsistent languages.

sentence or no sentence of the language is true (including atomic sentences, in which “tonk” does not occur).<sup>4</sup>

Such examples can be interpreted in diverse ways. Nevertheless, they show at least that to advance from understanding-assent links to understanding-truth links, let alone to understanding-knowledge links, is no trivial task.

One response to the examples is to stop trying to link understanding to knowledge and truth in this way, and try only to establish links to justification, conceived as non-factive. The hope would be to reach understanding-justification links like these:

- (UJl) Necessarily, whoever understands the sentence “Every vixen is a female fox” is justified in assenting to it.
- (UJt) Necessarily, whoever grasps the thought *Every vixen is a female fox* is justified in assenting to it.

But this retreat from knowledge and truth to justification does less than full justice to the examples. Imagine a dogmatic proponent of phlogiston theory, who continues to accept it long after the accumulating negative evidence has made this unjustifiable. Suppose that “phlogiston” does indeed provide a counterexample to the putative entailment from the understanding-assent link to the understanding-truth links. Thus although understanding a core of phlogiston theory necessitates assent to that core, because understanding the core necessitates understanding the term “phlogiston” and understanding “phlogiston” necessitates assent to the core of phlogiston theory, someone can understand the core despite its untruth. But if anyone can understand the core of phlogiston theory, its proponents can. Moreover, they do not stop understanding it when they unjustifiably refuse to take seriously the mounting negative evidence. Thus our last-ditch defender of phlogiston theory understands its core but is unjustified in assenting to it: the understanding-justification links fail too. For more blatantly defective concepts, the assent mandated by understanding-assent links may be unjustifiable from the start, as with “tonk.” In

<sup>4</sup> An example in which understanding is more clearly possible: Dummett (1973: 397, 454) claims that the rules for pejorative terms such as “Boche” suffer from a related kind of incoherence; Brandom (1994: 126; 2000: 69–70) and Boghossian (2003: 241–2), among others, have relied on his description of the practice of using such terms. I argue that it is mistaken in Williamson (2003a and 2008b), and suggest an alternative.

such cases too, an understanding-assent link which lacks the understanding-truth link also lacks the understanding-justification link.

Could one defend versions of UJt and UJl by qualifying the justification as *prima facie*? Consider someone who is introduced to a long list of mutually inconsistent theories of combustion, including phlogiston theory. Their content is explained without any assurance that there was ever any serious evidence for any of them. Irrationally, this person plumps for phlogiston theory and assents to its principles (unbeknownst to him, he is being influenced by happy associations from early childhood of the sound of the word “phlogiston”). By ordinary standards, he is linguistically competent with the sentences of phlogiston theory and grasps the corresponding thoughts, but he is not even *prima facie* justified in assenting to them, since he has no evidence, even by testimony, of their truth.

The examples do not motivate a retreat from knowledge and truth to non-factive justification. Rather, if they work, they show that some understanding-assent links have no positive epistemological upshot at all.

A different response to the examples is that they do not work: either the understanding-assent link fails or the understanding-truth link holds.

Since the relevant sentences or thoughts in the examples are clearly untrue, the understanding-truth link can hold in them only vacuously. That is, in such pathological cases, understanding is impossible: no meaning or concept is there to be grasped.<sup>5</sup> This response seems plausible for “tonk,” for any serious attempt to apply the “tonk” rules would lead to almost immediate disaster. The envisaged response also makes the links from understanding to truth and any positive epistemic status hold vacuously. Where there is no understanding, we can hardly expect much of a positive epistemological upshot from a constraint on understanding. A trickier question is whether such possibilities of an illusion of understanding have negative epistemological repercussions for cases of genuine understanding, since a skeptical doubt can arise for the subject in the latter cases too as to whether the understanding is not an illusion. If it could avoid such repercussions, this response might maintain a general entailment from understanding-assent links to understanding-knowledge links and the rest.

<sup>5</sup> See Peacocke (1992: 21) and Boghossian (2002).

However, the response is less plausible for “phlogiston” and some of the other examples than for “tonk,” since communities used the rules for “phlogiston” and “true” for years before running into any trouble.<sup>6</sup> There does seem to be some sort of difference between understanding the word “phlogiston” and not understanding it. Although speakers cannot know the reference of a term if it has none, they can attain some sort of ordinary linguistic competence with it, and in that attenuated sense understand it. If such understanding of theoretical terms requires understanding-assent links in general, it is unclear why it should fail to do so for the term “phlogiston” in particular. Similarly, even if sentences with “phlogiston” fail to express propositions, because “phlogiston” fails to refer, there is still an attenuated sense in which some speakers have the empty concept *phlogiston*, an empty mental vehicle, while others do not. If such possession of theoretical concepts requires understanding-assent links in general, it is unclear why it should fail to do so for the concept *phlogiston* in particular.

Alternatively, someone might maintain that the understanding-assent links in these examples fail, but that understanding-assent links for other sentences or thoughts hold; the examples involve genuine understanding. On this view, understanding-assent links may still be held to entail the corresponding understanding-knowledge links. It claims that the examples picked the wrong candidates for understanding-assent links. Either such links hold only for non-defective words or concepts or for those defective cases they hold only for cautiously circumscribed sentences or thoughts. For instance, rather than the core of phlogiston theory itself, we might have the conditional “If phlogiston exists then . . . ,” with that core filling in the dots. Arguably, however, since “phlogiston” fails to refer, that conditional too fails to express a proposition, so even this more cautious sentence is not true, although it is also not false. A more general objection is that this response treats our practices as though they are bound to have anticipated from the start all problems that could subsequently arise for them. Presumably, if understanding-assent links hold, they do so because they are built into the linguistic or conceptual practices at issue. Consider, for instance, the

<sup>6</sup> Boghossian (2003: 242–3), which represents a change of view from Boghossian (2002).

hypothesis that understanding “true” necessitates assent to a disquotational principle carefully and ingeniously modified to avoid all the semantic paradoxes. Since they scarcely ever arise in ordinary life, why was our ordinary practice with the word “true” tailored in advance to avoid them? Indeed, the puzzlement they cause suggests quite the opposite. That such precautions are part of every possible linguistic or conceptual practice is even less likely. If understanding-assent links hold for some other reason than that they are built into the linguistic or conceptual practices at issue, what is that other reason? Even if one moderates the approach by substituting understanding-justification links for understanding-knowledge links, a version of the objection still applies. If our linguistic or conceptual practices can make assent to inference rules a precondition of understanding, nothing seems to stop bad practices from requiring assent to rules, like those for “tonk,” that generate consequences not involving the original word or concept at issue. Such consequences may include arbitrary pernicious dogmas (such as racist ones) for which no justification is provided. More cautious fallbacks need not even implicitly have been provided; the practice simply breaks down once the dogma is abandoned. So this alternative way of maintaining a general entailment from understanding-assent links to understanding-justification links, let alone understanding-knowledge links, is unpromising. The objections tell equally against the putative understanding-knowledge or understanding-justification links, even if no attempt is made to *derive* them from understanding-assent links.

A more moderate response concedes that defective practices give rise to understanding-assent links without corresponding links to truth or any positive epistemological status, but maintains that understanding-assent links for non-defective practices do yield such links. For instance, one might try to tell a story on which understanding-assent links for non-defective practices constrain the reference of the relevant words or concepts so that the sentences or thoughts in the links come out true (for some defective practices, this constraint cannot be met). Under such conditions, understanding-assent links generate understanding-truth links. Thus assent to those sentences or thoughts (while understanding or grasping them) is, completely reliably, assent to truths. One might hope to squeeze understanding-knowledge links out of such reliability considerations, perhaps when

enhanced by an argument that the reliability is not completely hidden from the subject. Clearly, much work would be needed to vindicate such a programme.<sup>7</sup>

A lazy alternative simply postulates understanding-knowledge or understanding-justification links for non-defective practices without attempting to derive them from understanding-assent links. But this has little explanatory value. I understand “Every vixen is a female fox,” and it has some positive epistemic status for me. How does it get that status? How do I know “Every vixen is a female fox”? Why am I justified in assenting to it? The lazy theorist may try to dismiss the question, saying that it is simply part of our linguistic practice that “Every vixen is a female fox” has that positive epistemic status for whoever understands it. But the examples of defective practices show that it is not simply up to linguistic practices to distribute positive epistemic status as they please. That the practice is to treat a given sentence as having some positive epistemic status for competent speakers of the language does not imply that it really has that epistemic status for them. Their belief may be untrue and unjustified, however much the practice deems otherwise. Thus the only plausible way to make the relevant practice guarantee the putative link from understanding to the positive epistemic status is by making absence of the epistemic status constitute absence of understanding, just as absence of assent was supposed to do. On this account, whoever does not know “Every vixen is a female fox” or is not justified in assenting to it *thereby* fails to understand it. But this direction of explanation does not trivialize the positive epistemic status, to which it assigns the role of constituter, not constituted. Thus the lazy theorist cannot simply dismiss the question: how does “Every vixen is a female fox” gets its positive epistemic for whoever understands it? Positing direct links from understanding to knowledge or justification does not remove the need for substantive epistemology here. Indeed, it makes the armchair nature of understanding problematic. Even when the relevant sentence or thought has the positive epistemic status at issue, the reason is not simply that the linguistic or conceptual practice deems it to be so – which of course is not to say that the practice is

<sup>7</sup> The treatment of the issue in Boghossian (2003) is of this general kind. For detailed criticism see Williamson (2003a).

irrelevant to its epistemic status. In any case, if understanding-assent links fail, as is argued below, then *a fortiori* so do understanding-knowledge links, and understanding-justification links turn out to fail for similar reasons.

Let us consider understanding-assent links in more depth. If they hold, with or without normative consequences, they should cast some light on the actual practice of philosophy. For if an understanding-assent link holds for a philosophically significant sentence, and we do understand it, then we do assent to it, whether or not we are justified in doing so. But the next sections argue that understanding-assent links fail even for paradigms of “analyticity.” The main focus will be on the simplest cases, since those are the ones for which understanding-assent links have the best chance: if they fail there, they fail everywhere. We will start by examining unqualified understanding-assent links, beginning at the level of language. They fail. We then consider various ways of loosening them.

### 3

In their classic response to Quine’s critique of the analytic-synthetic distinction, Grice and Strawson give the sentence “My neighbor’s three-year-old child is an adult” as an example of a sentence that we could not understand someone using with its ordinary literal meaning to make an assertion (1956: 150–1). That suggests an understanding-assent link for the sentence “No three-year-old child is an adult”: necessarily, whoever understands it assents to it. But the link fails. Someone may believe that normal human beings attain physical and psychological maturity at the age of three, explaining away all the evidence to the contrary by *ad hoc* hypotheses or conspiracy theories (many three-year-olds pretend to be eighteen-year-olds in order to vote, the abnormally polluted local water slows development, and so on). However foolish those beliefs, they do not constitute linguistic incompetence. Friends of analyticity will reply that the example was badly chosen. It is therefore best to start with the most elementary examples possible.

Here is an elementary logical truth:

- (1) Every vixen is a vixen.

Few quantified logical truths are simpler than (1), in either syntactic complexity or the number of steps needed to derive them in a standard system of natural deduction rules.<sup>8</sup>

One may be tempted to endorse understanding-assent links for (1):

- (UAl') Necessarily, whoever understands the sentence “Every vixen is a vixen” assents to it.
- (UAt') Necessarily, whoever grasps the thought *every vixen is a vixen* assents to it.

Are UAl' and UAt' true? Consider two native speakers of English, Peter and Stephen.

Peter's first reaction to (1) is that it seems to presuppose:

- (2) There is at least one vixen.

On reflection, Peter comes to the considered view that the presupposition is a logical entailment. He regards the truth of “There is at least one F” as a necessary condition for the truth of “Every F is a G” quite generally, and the falsity of “There is at least one F” as a sufficient condition for the falsity of “Every F is a G”; he takes universal quantification to be existentially committing. More formally, he holds that “Every F is a G” is true if and only if (i) there is a value of the variable “*x*” for which “*x* is an F” is true and (ii) there is no value of the variable “*x*” for which “*x* is an F” is true while “*x* is a G” is not, and that “Every F is a G” is false if and only if it is not true. Of course, Peter does not always think in such theoretical, metalinguistic terms, but he resorts to them in rationalizing and defending his

<sup>8</sup> Parenthetical numerals such as “(1)” are taken throughout to refer to sentences rather than to thoughts. On a standard formalization of (1) as  $\forall x(Vx \rightarrow Vx)$ , one proves it by starting from an instance of the rule of assumption,  $Vx \vdash Vx$ , applying the standard introduction rule for  $\rightarrow$ , conditional proof, to discharge the premise, giving  $\vdash Vx \rightarrow Vx$ , followed by the standard introduction rule for  $\forall$ , universal generalization, to reach  $\vdash \forall x(Vx \rightarrow Vx)$  (no logical truth can be derived by the usual quantifier and structural rules alone, since none of them permits the discharge of all assumptions). A formalization of (1) closer to the English original uses a binary quantifier:  $\vdash (\text{EVERY}x(Vx; Vx))$  is derivable from  $Vx \vdash Vx$  in a single step by an appropriate introduction rule for **EVERY**.

pattern of assent and dissent to individual sentences. Peter also has the weird belief that (2) is false. For he spends far too much time surfing the Internet, and once came across a site devoted to propagating the view that there are no foxes, and therefore no vixens, and never have been: all the apparent evidence to the contrary has been planted by MI6, which even organizes widespread fox-hallucinations, so that people will protest about fox-hunting rather than the war in Iraq. Being a sucker for conspiracy theories, Peter accepted this one. Since he denies (2) and regards it as a logical consequence of (1), he also denies (1), and so does not assent to it.<sup>9</sup>

Stephen has no time for Peter's pet theories. What worries him is vagueness. He believes that borderline cases for vague terms constitute truth-value gaps. Like many truth-value gap theorists (such as Soames (1999)), he generalizes classical two-valued semantics by treating the gap as a third value ("indefinite") and using Kleene's three-valued "strong tables" (1952: 334), along the lines explained in Chapter 2. On Stephen's view, for "Every F is a G" to be true is for the conditional " $x$  is an F  $\rightarrow$   $x$  is a G" to be true for every value of the variable " $x$ "; for "Every F is a G" to be false is for " $x$  is an F  $\rightarrow$   $x$  is a G" to be false for some value of " $x$ ." On his semantics, for a conditional sentence with " $\rightarrow$ " to be true is for either its antecedent to be false or its consequent to be true, and for it to be false is for its antecedent to be true and its consequent false. Stephen also believes that some clearly female evolutionary ancestors of foxes are borderline cases for "fox" and therefore for "vixen." Consequently, for such an animal as the value of " $x$ ," " $x$  is a vixen" is neither true nor false, so the conditional " $x$  is a vixen  $\rightarrow$   $x$  is a vixen" is also neither true nor false, by the strong Kleene table for  $\rightarrow$ . Hence "Every vixen is a vixen" is not true; it is also not false, because the conditional is not false for any value of " $x$ ." Thus Stephen treats (1) as a truth-value gap. Of course, his initial reaction when presented with (1) is not to go through this explicit metalinguistic reasoning; he just says "What

<sup>9</sup> Alternatively, one can imagine that Peter thinks that foxes were only recently hunted to extinction, but that his presentist conception of time implies that (2) is true only if there is now at least one vixen. Yet another alternative is that Peter is a metaphysician who denies (2) on the grounds that putative macroscopic objects such as foxes do not exist, for if they did they would have vague boundaries, which are metaphysically impossible (compare Horgan (1998)).

about borderline cases?” But his refusal to assent to (1) as true is firm.<sup>10</sup>

We may assume that Peter and Stephen are wrong about (1), at least on its standard reading: it is in fact a logical truth. It is true however we interpret its only non-logical syntactically atomic constituent, “vixen,” given classical logic and two-valued semantics. If not, we can change the example, describing new characters who are deviant with respect to some sentence that really is an elementary logical truth. Peter and Stephen do not assent to (1). Thus, according to UAI’, Peter and Stephen do not understand (1) (with its standard English meaning). If so, they presumably misunderstand at least one of its constituent words or modes of combination. Is that the impression one would have in conversing with them?

Both Peter and Stephen treat “vixen” as synonymous with “female fox.” Stephen’s popular but mistaken theory of vagueness does not prevent him from understanding “vixen,” “female,” “fox” or their mode of combination. Even Peter’s conspiracy theory, however silly, involves no semantic deviation, just as religious fanatics who assert that there were never any dinosaurs do exactly that: they use the words “There were never any dinosaurs” to assert that there were never any dinosaurs, thereby expressing their belief that there were never any dinosaurs. Their problem is not that they misunderstand the word “dinosaur,” but that they have silly beliefs about evolution. Peter, like Stephen, understands the word “vixen.”

The best candidate for a word or mode of composition in (1) that Peter and Stephen misunderstand is “every.” Is it a good enough candidate? Peter’s not uncommon conception of the existential commitments of universal quantification makes little difference in practice, for when sentences of the form “Every F is a G” occur in conversation, “There is at least one F” tends to be common ground among the participants anyway. It is (usually, not always) a pragmatic presupposition in the sense of Stalnaker (1999). Pragmatically, Peter adjusts his conversation to a society that obstinately retains its belief in the existence of foxes much as members of many other small

<sup>10</sup> Note that while Peter assents to the conditional “If there are vixens, then every vixen is a vixen,” Stephen does not, because it has a true antecedent and an indefinite consequent, and is therefore itself indefinite on the Kleene semantics. Given the qualifications in Boghossian (2003), this makes Stephen more problematic than Peter for Boghossian’s program.

sects with unpopular beliefs have learned to adjust to an unenlightened world. Stephen's deviation is less localized than Peter's, because his Kleene-inspired semantics turns many universal generalizations with empirical predicates into truth-value gaps. In practice, however, he often manages to ignore the problem by focusing on a small domain of contextually relevant objects among which there are no borderline cases for the noun or complex phrase which complements "every." Occasionally he cannot avoid the problem and sounds pedantic, as many academics too, but that hardly constitutes a failure to understand the words at issue. When Peter and Stephen are challenged on their logical deviations, they defend themselves fluently. In fact, both have published widely read articles on the issues in leading refereed journals of philosophy, in English. They seem like most philosophers, thoroughly competent in their native language, a bit odd in some of their views.

Someone might insist that Peter and Stephen appear to be using the word "every" in its standard sense because they are really using it in senses very similar to, but not exactly the same as, the standard one. Indeed, it may be argued, their non-standard senses were explained above, since in each case a truth-conditional semantics for the relevant fragment of English was sketched on which (1) is not true, whereas by hypothesis (1) is true on the standard semantics of English. But matters are not so simple. Peter and Stephen are emphatic that they intend their words to be understood as words of our common language, with their standard English senses. They are not making unilateral declarations of linguistic independence. They use "every" and the other words in (1) as words of the public language. Each of them believes that his semantic theory is correct for English as spoken by others, not just by himself, and that if it turned out to be (heaven forbid!) incorrect for English as spoken by others, it would equally turn out to be incorrect for English as spoken by himself. Giving an incorrect theory of the meaning of a word is not the same as using the word with an idiosyncratic sense – linguists who work on the semantics of natural languages often do the former without doing the latter. Peter and Stephen's semantic beliefs about their own uses of "every" may be false, even if they sometimes rely on those beliefs in conscious processes of truth-evaluation. Indeed, we may assume that Peter and Stephen do not regard the elaborate articulations of truth-conditions and falsity-conditions for "Every F is a G" above as

capturing the way in which they or other English speakers conceptualize the meaning of “every,” which they regard as a semantically unstructured determiner for which a homophonic statement of meaning would be more faithful: even for us “Every F is a G” is not strictly synonymous with “There is no F that is not a G,” since the former does not contain negation. For Peter and Stephen, the more elaborate articulations are simply convenient records of important logical facts about “every.” Only in tricky cases do they resort to their non-standard semantic theories in evaluating non-metalinguistic claims such as (1) expresses. Their non-metalinguistic unorthodoxy as to when every F is a G is not ultimately derived by semantic descent from metalinguistic unorthodoxy as to when “Every F is a G” is true; rather, their metalinguistic unorthodoxy is ultimately derived by semantic ascent from their non-metalinguistic unorthodoxy.

Of course, the intention to use words with their normal public meanings does not guarantee success: it can fail in cases of sufficiently gross and extensive error. But that does not suggest that the intention is *irrelevant* to whether someone is using the words with those meanings. The intention is normally successful, in the absence of special defeating circumstances, just as the intention to use a proper name with the same reference as it has in the rest of the community is normally successful. The question is whether Peter and Stephen’s eccentricities are sufficiently gross and extensive to constitute defeating circumstances. By ordinary standards, they are not. Although they look gross enough when seen in isolation, they are compensated for by Peter and Stephen’s normality in other respects.

Peter and Stephen are native speakers who learned English in the normal way. They acquired their non-standard views as adults. At least before that, nothing in their use of English suggested semantic deviation. Surely they understood (1) and its constituent words and modes of construction with their ordinary meanings then. But the process by which they acquired their eccentricities did not involve forgetting their previous semantic understanding. For example, on their present understanding of (1), they have no difficulty in remembering why they used to assent to it. They were young and foolish then, with a tendency to accept claims on the basis of insufficient reflection. By ordinary standards, Peter and Stephen understand (1) perfectly well. Although their rejection of (1) might on first acquaintance give an observer a defeasible reason to deny that they under-

stood it, any such reason is defeated by closer observation of them. They genuinely doubt that every vixen is a vixen. Nor are Peter and Stephen marginal cases of understanding: their linguistic competence is far more secure than that of young children or native speakers of other languages who are in the process of learning English. They joined the club of “every”-users; since they haven’t resigned or been expelled, they are still members.

If some participants in a debate have an imperfect linguistic understanding of one of the key words with which it is conducted, they need to have its meaning explained to them before the debate can properly continue. But to stop our logical debate with Peter and Stephen in order to explain to them what the word “every” means in English would be irrelevant and gratuitously patronizing. We cannot understand them better if we translate their word “every” by some non-homophonic expression, or treat it as untranslatable. The understanding they lack is logical, is not semantic. Their attitudes to (1) manifest only some deviant patterns of belief. Since there clearly could have been, and perhaps are, people such as Peter and Stephen, we have counterexamples to UAI’.

The argument that Peter and Stephen mean what we mean by their words exemplifies two interlocking themes: Quine’s epistemological holism, on which the epistemological status of a belief constitutively depends on its position in the believer’s whole system of beliefs, and Putnam and Burge’s semantic externalism (discussed in more detail below), on which the content of a belief constitutively depends on the believer’s position in a society of believers. Epistemological holism explains how unorthodoxy on one point can be compensated for by orthodoxy on many others, so that overall Peter and Stephen’s usage of the key terms is not beyond the pale of social acceptability; since they remain participants in the relevant linguistic practice, semantic externalism then explains how they can still use the terms with their normal public senses. But neither epistemological holism nor semantic externalism figured as *premises* of the argument. Rather, the argument appealed to features of the relevant systems of belief that make epistemological holism plausible, and to features of our ascription of beliefs that make semantic externalism plausible.

To try to save UAI’ by restricting it to rational agents would be pointless. By ordinary standards, Peter and Stephen are rational agents. Although they fall short of some high standards of rationality,

so do most humans. Understanding-assent links that do not apply to most humans would be of limited epistemological interest. The picture was that those who appear to reject analytic sentences can be excluded from the discussion because they lack the linguistic competence to engage in it; but we cannot exclude humans who reject such sentences on those grounds if the connection between rejecting them and lacking competence holds only for super-humans, not for humans.

The problem for UAI' is clearly not specific to sentences of the form “Every F is an F” Let us see how it generalizes to rules of inference.

It is often claimed that assent to arguments by modus ponens of the form “If A then B; A; therefore B” is a precondition for understanding the word “if” (Boghossian 2003, for instance). Indeed, this is a standard example in the literature. However, Vann McGee, a distinguished logician, has published purported counterexamples to modus ponens for the indicative conditional in English. Here is one of them; the others are similar:

Opinion polls taken just before the 1980 election showed the Republican Ronald Reagan decisively ahead of the Democrat Jimmy Carter, with the other Republican in the race, John Anderson, a distant third. Those apprised of the poll results believed, with good reason:

If a Republican wins the election, then if it's not Reagan who wins it will be Anderson.

A Republican will win the race.

Yet they did not have reason to believe:

If it's not Reagan who wins, it will be Anderson. (McGee 1985: 462)

With reasonable confidence, they combined assent to both premises of an argument by modus ponens with dissent from the conclusion, so they rejected the argument.<sup>11</sup> If McGee's examples are counterexamples to modus ponens, they are also counterexamples to the claim that assent to instances of modus ponens is necessary for understanding “if.” But let us assume, with the majority, that modus ponens is

<sup>11</sup> The formulation in the text is intended to distinguish the case from examples in which speakers' confidence in each premise of a modus ponens argument is just above a probabilistic threshold which their confidence in the conclusion is just below. In McGee's case, speakers are sufficiently confident of the conjunction of the two premises.

valid, so McGee's examples are not in fact counterexamples.<sup>12</sup> Perhaps the conclusion was true, because Reagan won; although the poll was not misleading, our usual methods for evaluating conditionals lead us astray in this case. A currently popular objection to the examples is that they depend on an illicit shift of context, perhaps in the treatment of "If it's not Reagan who wins, it will be Anderson" between the consequent of the first premise and the conclusion.<sup>13</sup> But even if some such confusion *causes* the pattern of assent and dissent to the premises and conclusion, the *effect* is that McGee and his envisaged speakers end up accepting the premises and rejecting the conclusion in a single context, when they look back on all three sentences.<sup>14</sup> They genuinely reject a genuine instance of modus ponens.<sup>15</sup> Such reactions do not manifest the superimposition of a perverse semantic or logical theory on native speaker intuitions; they flow from native speaker intuitions themselves in a fairly natural way, despite being mistaken.

<sup>12</sup> For early critical reactions to McGee's examples see Sinnott-Armstrong, Moor, and Fogelin (1986), Lowe (1987) and Over (1987). But some authors have accepted the examples (Lycan 2001: 66–7).

<sup>13</sup> Recent examples of context-shifting charges include Nolan (2003: 264) and Gauker (2005: 86).

<sup>14</sup> Contrast McGee's example with instances of modus ponens such as "I know that I have hands; if I know that I have hands then I know that I'm not a brain in a vat; therefore, I know that I'm not a brain in a vat." Many people accept the premises and reject the conclusion when they encounter them in that order. However, once they have rejected the conclusion, they are typically inclined to retract their acceptance of the first premise, not out of concern for modus ponens but because it no longer looks plausible to them in its own right, in the new context that arises once the skeptical possibility becomes relevant. For contextualists in epistemology, this is a paradigm case of context-shifting (Stine (1976), Cohen (1988), DeRose (1995), Lewis (1996); see Hawthorne (2004), Stanley (2005) and Williamson (2005b) for some critical discussion and more references). By contrast, the premises of McGee's argument continue to look plausible to those who reject the conclusion.

<sup>15</sup> Edgington (2001: 408) suggests that McGee's example is not a genuine instance of modus ponens on the grounds that the first premise has a misleading surface form; on her view, conditionals do not express propositions, so what look like conditionals with conditional antecedents or consequents must be reinterpreted. It is doubtful that such a view is consistent with a systematic account of the structure of English sentences, which permits a wide variety of such embeddings, for example, "If it is the case that if it's not Reagan who wins it will be Anderson, then a Republican will win the race."

Does McGee not understand the English word “if”? In conversation, he appears to understand it perfectly well. By ordinary standards, he *does* understand it. Before he had theoretical doubts about modus ponens, he understood the word “if” if anyone has ever understood it. Surely his theoretical doubts did not make him cease to remember what it means. Moreover, his doubts derive from taking at face value a natural pattern of native speaker reactions to an ingeniously chosen case. If he counts as not understanding “if,” so do millions of other native speakers of English.

Could we invoke the division of linguistic labor (Putnam 1975: 228), and say that making any given inference by modus ponens is a precondition only for *full* understanding of “if,” the kind of understanding characteristic of the expert rather than the layman? The trouble is that McGee *is* an expert on conditionals. He publishes on them in the best journals. He does not defer in his use of “if” to any higher authorities. He may lack some theoretical understanding of conditionals, just as experts on neutrinos may lack some theoretical understanding of neutrinos, but none of that amounts to any lack of linguistic competence with “if” or “neutrino” at all.

Are only some arguments by modus ponens such that assent to them is a precondition for understanding “if”? Presumably, McGee will accept most arguments by modus ponens. However, any particular such argument might be rejected by another expert on conditionals, on the basis of a subtle theoretical argument. By hypothesis, the expert would be mistaken, but making a subtle theoretical error does not constitute linguistic incompetence.

The problem is not just the vagueness of natural languages. Similar problems arise for carefully constructed formal languages. Consider modus ponens for the material conditional  $\rightarrow$ , explained by the standard truth-table. It is equivalent to disjunctive syllogism: from  $A$  and  $\neg A \vee B$  derive  $B$ . Technically competent relevance logicians and dialetheists such as Graham Priest reject disjunctive syllogism (Priest 1995: 5). According to him, the best account of paradoxes such as the Liar is that in special circumstances a sentence can be both true and false; one can be on different lines of the truth-table simultaneously. When  $A$  is true and false while  $B$  is merely false, the premises of disjunctive syllogism are true (for  $A$  is true; since  $A$  is also false,  $\neg A$  is true, so  $\neg A \vee B$  is true), while its conclusion is straightforwardly false. Whatever the errors underlying the rejection of modus

ponens for  $\rightarrow$ , they do not arise from a lack of linguistic competence with  $\rightarrow$  on the part of relevance logicians and dialetheists.

As a final example, consider the natural deduction rules for conjunction. Instances of the introduction rule are arguments of the form “A; B; therefore A and B.” Instances of the elimination rule are arguments of the converse forms “A and B; therefore A” and “A and B; therefore B.” These are just about the simplest rules for a non-trivial binary connective. One must formulate what acceptance of the introduction rule requires with particular care, since the probability of a conjunction may be less than the probability of either conjunct. Iterations of the introduction rule yield the Lottery and Preface paradoxes. Given a lottery known to have at most a million tickets and only one winner, each premise of the form “Ticket  $i$  will lose” is overwhelmingly probable, even though their conjunction is known to be false. The author of a book may endorse each individual statement in it, yet admit in the preface that, despite all her efforts, it is bound to contain errors, and on those grounds reject the conjunction of the individual premises. Of course, these paradoxes do not show that the introduction rule fails to preserve truth, although they might be used as grounds for rejecting the rule by a theorist who (mistakenly) used a probabilistic criterion for acceptance. The elimination rule does not suffer from these problems, since the probability of a conjunction is never higher than the probability of any given conjunct.

Let us therefore concentrate on the elimination rule for conjunction, as having the best chance of being non-discretionary for competent speakers.<sup>16</sup> Consider Simon, whose view of vagueness resembles Stephen’s, except that Simon’s practice conforms to a semantics with Kleene’s weak three-valued tables rather than his strong ones. On these tables, a conjunction is indefinite (neither true nor false) if at least one conjunct is, irrespective of the value of the other conjunct; the same principle is applied to disjunction, the material conditional and negation (Kleene 1952: 334). Furthermore, Simon regards both

<sup>16</sup> In discussion, Boghossian suggested conjunction elimination as a fallback example of a non-discretionary rule if modus ponens fails. Peacocke writes of the possession-condition for the concept of conjunction, “On any theory, this possession-condition will entail that thinkers must find the transition from A and B to A compelling, and must do so without relying on any background information” (2004: 172).

truth and indefiniteness as designated (acceptable) semantic values for an assertion: what matters is to avoid falsity. In a borderline case, some speakers say “Jack is bald,” others with equal vehemence say “Jack is not bald”; they may persist even when they recognize that the dispute cannot be resolved. According to Simon, both assertions are acceptable. In answer to the question “Is Jack bald?,” even the answer “He is and he isn’t” is acceptable. Although Simon does not assign the value “T” to “Jack is bald,” that metalinguistic reservation is consistent with assenting to the sentence, that is, with believing that Jack is bald under the guise of that very sentence (similarly, supervaluationists about vagueness reject the disquotational inference from “‘Jack is bald’ is not true” to “Jack is not bald”). The joint implication of Simon’s principles is that any complex sentence formed by the application of the specified operators to simpler sentences, at least one of which is borderline, has a designated value – of course, on Simon’s view, most such sentences should not be uttered, on the pragmatic grounds that they violate the conversational maxim of relevance (Grice 1989: 27). Suppose that “A” is simply false while “B” is borderline. Consequently, for Simon, “B” is indefinite, so “A and B” is also indefinite. Thus the corresponding instance of conjunction elimination – “A and B; therefore A” – has a designated premise and an undesignated conclusion. On these grounds, Simon rejects the conclusion of that instance while accepting its premise (although he points out that asserting the premise would be pragmatically misleading in most contexts, since “B” is irrelevant to its status). In other cases, he treats the premise merely as a supposition, but still rejects the deduction from it to the conclusion. Once again, this need not reflect incompetence with the English language. Conjunction elimination is no exception to the general pattern. Arguably, violations of conjunction elimination are actual, not just possible, in the Conjunction Fallacy, a much-studied, widespread and robust psychological phenomenon in which subjects assign a higher probability to a conjunction than to one of its conjuncts.<sup>17</sup>

<sup>17</sup> The seminal paper is Tversky and Kahneman (1983). See also Kahneman and Frederick (2002), Sides, Osherson, Bonini, and Viale (2002) and Jönsson and Hampton (2006). We can also imagine speakers who reject instances of conjunction elimination through muddling truth and conversational appropriateness. “Did she take the money and give it back? Yes. Did she take the money? No, she took-the-money-and-gave-it-back.”

No given argument or statement is immune from rejection by a linguistically competent speaker. Quine's epistemological holism in "Two Dogmas" undermines his notorious later claim about the deviant logician's predicament: "when he tries to deny the doctrine he only changes the subject" (1970: 81).

Understanding words in a natural language has much to do with the ability to use them in ways that facilitate smooth and fruitful interaction with other members of the community. That ability can be realized in indefinitely various forms. Speakers can compensate for their deviance on one point by their orthodoxy on others, their ability to predict the reactions of non-deviant speakers, their willingness in the long run to have their utterances evaluated by public standards. As we have seen, such compensation is often possible when the deviance results from localized interference in the normal practice of using a word by high-level theoretical concerns. Thus there is no litmus test for understanding. Whatever local test is proposed, someone could fail it and still do well enough elsewhere with the word to count as understanding it. Could an inferentialist reply that such objections trade on a loose everyday sense of "understanding" that must be replaced by something more precise for theoretical purposes? It is far from clear that a stricter sense would do a better job. The relevant features of the ordinary conception of understanding are not mere unreflective sloppiness. Rather, they are an appropriate response to an important constraint on a theory of linguistic meanings: that there is little point in talking about them unless they can be shared across significant differences in belief, between different individuals at the same time or the same individual at different times. They can survive factual learning and factual disagreement. Although inferentialist accounts respect the letter of that constraint, they violate its underlying spirit, by setting inflexible limits to the scope for genuine disagreement. The more holistic ordinary notion of understanding permits localized disagreement at virtually any point.

Cases of logical deviance hint at ways in which the failure of individualist accounts of meaning go deeper than the immediate lessons of the original anti-individualist arguments of Putnam (1975) and Burge (1979). Their cases are often analyzed in terms of a distinction between experts with full understanding and lay-people with partial understanding who defer to the experts, in virtue of which one may

correctly ascribe to them attitudes to the contents that experts determine.<sup>18</sup> Such asymmetries are postulated by Putnam's Hypothesis of the Universality of the Division of Linguistic Labor:

Every linguistic community . . . possesses at least some terms whose associated "criteria" are known only to a subset of the speakers who acquire the terms, and whose use by the other speakers depends upon a structured cooperation between them and the speakers in the relevant subsets. (Putnam 1975: 228)

But, as we have seen, experts themselves can make deviant applications of words as a result of theoretical errors and still count as fully understanding their words. Although they defer to nobody on the matters at issue, they are more than adequately integrated members of the speech community with respect to those very words. Their assignments of meaning to those words are not parasitic on the assignments that more privileged individuals make. Rather, each individual uses words as words of a public language; their meanings are constitutively determined not individually but socially, through the spectrum of linguistic activity across the community as a whole. The social determination of meaning requires nothing like an exact match in use between different individuals; it requires only enough connection in use between them to form a social practice. Full participation in that practice constitutes full understanding. That is why there is no litmus test for understanding.<sup>19</sup>

<sup>18</sup> An example is Peacocke's discussion of deference-dependent propositional attitude ascriptions (1992: 29–33). Burge (1986) extends his earlier arguments in ways related to the arguments of this chapter, in his account of the understanding of words such as "sofa," and argues for such a deeper lesson. Goldberg (2000) replies on behalf of Burge to Bach (1988) and Elugardo (1993).

<sup>19</sup> For a related conclusion concerning lexical competence in a shared language see Marconi (1997: 56). For the relevance of the model of full understanding as full induction into a practice to the theory of vagueness see Williamson (1994a: 211–12). It is not implied that no similar issue could arise for understanding on the part of a single isolated individual, for such an individual's meanings and concepts are constitutively determined, at least in part, by their dispositions over a range of counterfactual circumstances; those dispositions and their bearings may be hard to survey from the limited standpoint of the actual circumstances.

## 4

Peter and Stephen understand (1) without assenting to it; UAI' fails. Someone sympathetic to the spirit of understanding-assent links might concede that much while arguing that its upshot is only a superficial loosening of those links. If the deviance results only from erroneous theorizing that overlays an ordinary understanding of the terms, may not the links still hold at the underlying level?

However, we have already seen reason to doubt that deviance can only arise from theorizing extrinsic to speakers' ordinary understanding of the words. Vann McGee's examples exert an intuitive pull on native speakers, irrespective of and even contrary to their theoretical predilections. We can also imagine untheoretical native speakers whose unreflective patterns of assent and dissent to non-metalinguistic sentences are those which Peter, Stephen, and Simon respectively recommend, although they lack the reflective capacity to rationalize those patterns by appeal to formal semantic theories. They too would be able to fit in well enough with the rest of the linguistic community, to engage smoothly in useful communication and adjust to their differences with other speakers in order not to attract too much attention. They too would use their words as words of the public language, rather than declaring unilateral linguistic independence. How do we know that there are not in fact many such native speakers of English around us? Once we concede that Peter, Stephen, and Simon are competent speakers, we can hardly refuse the same classification to other speakers merely on grounds of their unacquaintance with formal semantics.

What might be claimed, in the case of both theoretical and untheoretical deviant native speakers, is that the deviance is some kind of performance error which leaves their underlying competence intact: at some basic level they have the required dispositions, which they fail to manifest as a result of interfering factors, such as computational limitations, conflicting dispositions to take cheap and dirty intellectual short-cuts, and so on. On this view, Peter and Stephen still have a disposition to assent to (1), masked by their later theorizing; they use "every" and other words and modes of construction with the same senses as the rest of us because they have the same underlying

inferential dispositions as the rest of us.<sup>20</sup> At some deep level, they have a disposition to accept (1) as true. That disposition is prevented from manifesting itself by conscious reflection at an overlying level of theory-construction, just as someone's pet views about grammar might interfere with their performance in speech while having no effect on the syntactic competence which they possess in virtue of their underlying linguistic competence. For untheoretical speakers, the interfering factors are unconscious, but the effect is similar. UAl' and UAt' might therefore be watered down as follows:

- (UDAl') Necessarily, whoever understands the sentence "Every vixen is a vixen" has a disposition to assent to it.
- (UDAt') Necessarily, whoever grasps the thought *every vixen is a vixen* has a disposition to assent to it.

Having a disposition to assent does not entail assenting. Thus UDAI' and UDAt' are consistent with the denials of UAl' and UAt'. Do Peter and Stephen have the disposition to assent to (1) despite happening not to assent to it? If understanding is linked to such dispositions to assent in these cases, one might even try to use that to explain how it is also linked to dispositions to know, along lines similar to those sketched in Section 2. But are UDAI' and UDAt' true?

There are two salient ways to fill out the dispositional story: at the *personal level* or the *sub-personal level*. At the personal level, the postulated dispositions require something like counterfactual conditionals to the effect that sufficient conscious reflection and exposure to further arguments would bring the person to assent. Thus Peter and Stephen would assent to (1) if only they thought about it more and talked to more experts. By contrast, at the sub-personal level, the postulated dispositions are grounded in something like an unconscious reasoning module, even if the personal-level counterfactual conditionals are false. Thus the default outcome of Peter and Stephen's underlying competence is assent to (1), even if stable dispositions from other sources irreversibly override that default.

<sup>20</sup> Eklund (2002: 262) defends such a view of logical deviance. See Martin (1994), Lewis (1997), Martin and Heil (1998), Bird (1998), and Mumford (1998) for some basic issues about masked dispositions. Harman (1999: 213) relies on defeasible inferential dispositions in his conceptual role semantics.

An analogous contrast arises for syntax. As a standard example, native speakers of English tend to reject (3) at first sight as ill-formed:

(3) The horse raced past the barn fell.

They want to insert “and” between “barn” and “fell.” But they tend to change their minds about (3) when asked to consider the result of inserting “that was” between “horse” and “raced” instead: they realize that the original string was well-formed after all; “the horse” is the object, not the subject, of “raced.” Conversely, native speakers often unreflectively accept ill-formed strings as well-formed, for example when a plural verb is separated from its singular subject by a long intervening string that includes a plural noun, but can be brought to acknowledge their mistake, as when a draft is corrected. On a personal level account, such conscious reflective judgments, actual or counterfactual, are constitutive of well-formedness. On the contrasting sub-personal level account, those judgments play a merely evidential role: what constitutes well-formedness is the structure of the syntactic component of the unconscious language module, even if the person’s conscious reflective judgment is irreversibly contrary as a result of extraneous factors, such as their dogmatic commitment to a pet theory of syntax.

The personal level account fails to shield UDAl' and UDAt' from the counterexamples of Peter and Stephen. For, by hypothesis, their refusal to assent to (1) is stable under conscious reflection, exposure to further arguments and so on. Like many people, not least philosophers, they are obstinate in defense of their favorite views, willing to make whatever *ad hoc* moves are needed to retain them. One knows in advance that the task of dissuading them is hopeless, however good one’s objections: a common experience in philosophy. As Peter and Stephen became comfortable with their deviant theories they gradually ceased to feel even an initial temptation to assent to (1), we may assume, although they still remember what it was like to feel such a temptation. They assimilate the change to one in which education gradually eradicates the tendency to make a particular false assumption. Perhaps years of browbeating or social ostracism would cause them to change their minds, but that applies to almost any belief; it is poor evidence that an underlying disposition to assent was present all

along. Would Peter and Stephen assent to (1) if they lacked their conscious theoretical commitments? Perhaps not, but that counterfactual would show little. The possibility of untheoretical analogues of Peter and Stephen has already been raised. They lack the conscious theoretical commitments but still do not assent to (1). If it is objected that the untheoretical analogues, unlike Peter and Stephen, do not understand (1) with its normal English sense because they lack the required unconscious cognitive structures, that is in effect to switch to the sub-personal version of the dispositional account. On the personal level account, Peter and Stephen are *not* disposed to assent to (1). If that makes them irrationally obstinate, they are no more so than many philosophers and non-philosophers in defense of a favorite view.

The sub-personal level story has more room for maneuver in defense of UDAI' and UDA $\ddot{a}$ '. It can insist that although Peter and Stephen's personal refusal to assent to (1) is stable under conscious reflection and exposure to further arguments, they retain a disposition to assent to (1) in virtue of features of their unconscious logic rules. This requires the postulated rules to be encased in some sort of psychological module, for if they consisted only in general habits of reasoning, Peter and Stephen's earlier habits could eventually be erased by their later ones, and the disposition to assent to (1) would disappear. The module must include rules for deduction, since that is the kind of reasoning relevant to (1). This module may be a component of an overall semantic module (after all, we are considering (1) as a candidate for analyticity). If the grounds for assent to (1) were merely inductive – that we have never observed a vixen that was not a vixen – people who understood (1) could reasonably refuse to assent to it on the grounds that they had observed too few vixens to be in a position to judge. A *prima facie* attractive conjecture is that the deductive rules would include analogues for natural language connectives of the introduction and elimination rules in a Gentzen-style system of natural deduction. But do humans have a module that includes unconscious logic rules of the required sort?

One might suppose the primary adaptive value of a cognitive module to be its capacity to perform a specific type of useful information processing quickly and reliably enough for the purposes of action in a changing environment. Its design can exploit special features of the type of task to which it is dedicated, in order to achieve efficiencies that would be impossible for a general purpose central processing

unit. A diversion through higher mental processes, in particular through consciousness, would be slower and less reliable. Thus one might expect unconscious modular deductive reasoning to pay its way by the speed and reliability of its results, just as modules for vision and natural language processing seem to do. Naturally, performance would tail off as the complexity of problems increased, but there should be good performance over a worthwhile range of non-trivial problems. Is that prediction borne out?

Evidence from empirical psychology, amassed over several decades, suggests that most humans are strikingly bad at even elementary deductive reasoning, a finding which should not surprise those who have taught introductory logic. For example, in the combined results of over 65 large-scale experiments by different researchers on simple conditional reasoning, although 97 percent (not 100 percent!) of subjects endorsed modus ponens, only 72 percent endorsed modus tollens (if A then B; not B; therefore not A), while as many as 63 percent endorsed the fallacy of affirming the consequent (if A then B; B; therefore A) and 55 percent endorsed the fallacy of denying the antecedent (if A then B; not A; therefore not B). When the antecedent is negative, affirming the consequent overtakes modus tollens in popularity.<sup>21</sup> In some cases, when a further premise of the form “If C then B” is added to modus ponens only a minority endorses the inference (Byrne 1989).<sup>22</sup> Similar phenomena arise for elementary syllogistic reasoning.

Performance greatly improves when the conditional premise in a reasoning task has a realistic deontic content, such as “If you use a second class stamp, then you must leave the envelope unsealed” (Manktelow and Over 1987, Wason and Shapiro 1971). In general, the real-life credibility or otherwise of premises and conclusion strongly influences judgments of validity and invalidity.

<sup>21</sup> See Schroyens and Schaeken (2003); the percentages are as summarized by Oaksford (2005: 427).

<sup>22</sup> Is it still modus ponens if there is an extra redundant premise? If not, then humans apply modus ponens only in the most artificial circumstances, since in practice we always have further information. Moreover, people without formal education tend to do *worst* in reasoning tasks with artificial premises from which all background information has been screened out (see Harris (2000: 94–117) for discussion). Such a restriction would make a disposition to assent to modus ponens a rather artificial test for understanding “if.”

For simple problems in formal deductive reasoning, when the specific subject matter provides no helpful clues, success is significantly correlated with intelligence, in whatever sense it is measured by IQ tests, SAT scores or the like (Stanovich and West 2000). For some simple tasks, success is rare except among those with the intelligence of able undergraduates (Newstead *et al.* 2004; the samples in the experimental literature tend to consist of university students, since they are the most easily available subjects). Contrast this with the efficient success which humans typically show in judging whether short strings of words constitute well-formed sentences of their native language, for example. There is little sign of anything modular that contains formal rules to subserve conscious deduction, whether conceived as part of a language module or as part of a reasoning module.

Of course, there may be sub-personal processes whose inner workings can conveniently be represented as employing deductive rules, just as there may be sub-personal processes whose inner workings can conveniently be represented as employing differential equations, for example to process perceptual input, in even the most mathematically ignorant subjects. But that is not quite the issue. We are questioning the existence of a sub-personal basis for an unmanifested disposition to assent, that is, to perform an action at the personal level. The problem is that the data of normal performance tell against the hypothesis of a set of deductive rules (semantic or not) unconsciously employed as the primary route to conscious assent in the relevant normal cases.

A widespread, although not universal, view among psychologists of reasoning is that humans have two reasoning systems. In the terminology of Stanovich and West, System 1 is associative, holistic, automatic, relatively undemanding of cognitive capacity, relatively fast, and acquired through biology, exposure, and personal experience; its construal of reasoning tasks is highly sensitive to personal, conversational, and social context. System 2 is rule-based, analytic, controlled, demanding of cognitive capacity, relatively slow, and acquired by cultural and formal tuition; its construal of reasoning tasks is rather insensitive to personal, conversational, and social context.<sup>23</sup> System 1 lacks the formal rules that enable deductive rea-

<sup>23</sup> See Stanovich and West (2000: 659), where a list is also provided of earlier authors who have proposed similar views.

soning to succeed in the absence of helpful clues from the content of premises and conclusion. Although defeasible and only moderately reliable, it performs an important role in tasks of the kind for which it presumably evolved, such as integrating new information from perception or testimony with standing beliefs. System 1 is not a system for formal deductive reasoning. A suitably educated, highly intelligent person can achieve success in formal deductive reasoning by means of System 2, but it is not sealed off in an unconscious module.

How does this picture apply to Peter and Stephen? With respect to System 1, they fall within the normal range of human variation. They are slightly unusual with respect to System 2, which is in any case much more sensitive than System 1 to specific features of the individual's intelligence and education. But neither high intelligence nor a good education is needed to understand simple sentences like (1). Any System 2 differences at issue between Peter or Stephen and average speakers of English are wholly consistent with Peter and Stephen's competence in their native language. If Peter and Stephen do have any underlying disposition to accept (1) as true, it concerns their System 1. But aversion to universal generalizations with empty subject terms or borderline cases seems to be within the normal range of System 1 reasoning among native speakers. On the two systems picture, there is no reason to assume that all linguistically competent speakers have an underlying disposition to assent to (1).

The two systems picture has not been conclusively established; it may turn out to need modification. Nevertheless, it throws into relief the empirical speculations on which the sub-personal understanding-disposition-to-assent links depend, and their clash with much current thinking in the psychology of reasoning. If the two systems picture is right to even a first approximation, the sub-personal links are in trouble.

How can System 1 or any other system evaluate deductive arguments without using formal rules for reasoning with logical constants in natural language, even if their effect is almost swamped by associations, heuristics, and other pragmatic factors?<sup>24</sup> There are alternatives. For example, one of the main psychological theories of deductive

<sup>24</sup> For such an approach see Braine and O'Brien (1991), criticized by Evans and Over (2004: 56–9).

reasoning is currently the *mental models* approach. Two of its leading proponents write:

The evidence suggests that it [the reasoning mechanism] is *not* equipped with logical rules of inference, which it sometimes uses correctly and sometimes misuses, misapplies or forgets. This analogy with grammar, which has seduced so many theorists, is a mistake. The reasoning mechanism constructs a mental model of the premises, formulates a putative conclusion, and tests its validity by searching for alternative models in which it is false. The search is constrained by the meta-principle that the conclusion is valid only if there are no such models, but it is not governed by any systematic or comprehensive principles. (Johnson-Laird and Byrne 1993: 178)

Thus subjects may erroneously classify an invalid argument as valid, because the unrepresentative sample of models they have examined includes no counter-model, and they wrongly treat it as representative. They may erroneously classify a valid argument as invalid, because they leave the process of constructing a counter-model incomplete, under the misapprehension that there is no obstacle to completing it. Background beliefs about the specific subject matter of an argument influence its classification because they influence which mental models are constructed. Johnson-Laird and Byrne argue that their theory gives the best fit to the empirical data.

On the mental models approach, the nearest one normally comes to employing deductive rules of inference is in the procedures for evaluating sentences (premises or conclusions) with respect to a given model, itself conceived as a mental representation.<sup>25</sup> But that process does not involve deductive reasoning in a natural language. Nor would natural deduction rules for the natural language connectives be very relevant; it is more like the construction of a truth-table. For example, in calculating the truth-value of a conditional in a model, one does not apply the rule of conditional proof to that very conditional if one already has the rules for constructing truth-tables.<sup>26</sup>

<sup>25</sup> Mental models need not be visualized (Johnson-Laird and Byrne 1993: 182). Johnson-Laird and Byrne also claim that human reasoning is a semantic rather than a syntactic process (*ibid.*: 180), but the significance of this claim is not entirely clear, since they treat reasoning as a manipulation of representations.

<sup>26</sup> Standard proofs of formalizations of (1) use conditional proof.

Evaluating a sentence in a model might involve something closer to an imaginative analogue of the processes that issue in complex perceptual judgments such as “Everybody over there is wearing a hat.” Not all such universally quantified conclusions are reached by deduction from further premises. One might employ this argument:

A is wearing a hat.  
B is wearing a hat.  
C is wearing a hat.  
Everybody over there is A, B, or C.  
Therefore:  
Everybody over there is wearing a hat.

But of course the final premise “Everybody over there is A, B, or C” is itself a universally quantified perceptual judgment. To suppose that it too was reached as the conclusion of a deductive argument is to start a futile regress.

Although the mental models theory does not apply to all human reasoning – for example, to the System 2 kind some humans learn to carry out in logic classes – it may apply to a high proportion of it. The theory is a salutary reminder that reasoning with logical constants need not be formal deductive reasoning, and that the empirical evidence suggests that in humans it usually is not.

One remaining concern is that logical skills must play some role in linguistic competence because logical features play a role in determining well-formedness. An example is the category of negative polarity items. Consider these sentences:

(4) If she ate any of the cake, she was hungry.  
(5)\* If she was hungry, she ate any of the cake.

“Any” is a negative polarity item. To a first approximation, the reason why “she ate any of the cake” is acceptable as the antecedent of the conditional but not as the consequent is that the antecedent is in a downward entailing (negative) context while the consequent is instead in an upward entailing (positive) context. A context C is upward entailing just in case whenever A entails B, C(A) entails C(B); C is downward entailing just in case whenever A entails B, C(B) entails C(A). Thus recognition of the logical features of contexts

seems to be needed in order to distinguish between well-formed and ill-formed sentences. But things are not so simple. Consider these sentences:

- (6) Exactly four people in the room were of any help.
- (7) Few people in the room were of any help.

Logically, “few” creates a downward entailing context; “exactly four” does not. However, (6) is acceptable provided that in the context it is taken to imply (7), but not generally otherwise. Thus the phenomenon involves a significant pragmatic element: which contexts are suitable for “any” cannot be determined on purely logico-linguistic grounds. If we disagree with the speaker of (6) about how many people were in the room or what proportion of them could have been expected to help, we may find her use of “any” inappropriate without regarding her as *linguistically* incompetent. Similarly, if a speaker has deviant views as to which contexts are downward entailing, but uses “any” in just those contexts that she treats as downward entailing, we might find her deviant use of “any” inappropriate without regarding her as linguistically incompetent, precisely because the deviation in use is explained by logical rather than linguistic unorthodoxy. Thus the role of logical knowledge in such cases does not make it part of purely linguistic competence. All our knowledge is potentially relevant to judging the appropriateness of a given use of “any.”<sup>27</sup>

Suppose, nevertheless, that our classification of strings such as (4)–(7) as well- or ill-formed does depend on some prior classification of contexts as downward entailing or not. The question remains: is that classification available for unconscious reasoning that would issue in conscious assent to supposedly analytic sentences? To identify

<sup>27</sup> Ladusaw (1996: 325–37) surveys issues concerning negative polarity. Strictly speaking, the context of the antecedent of a counterfactual conditional is not downward entailing on standard logics of such conditionals, according to which strengthening of the antecedent fails; for example, although “It rained hard” entails “It rained,” “If it had rained, it would not have rained hard” does not entail “If it had rained hard, it would not have rained hard.” Nevertheless, negative polarity items are felicitous in the antecedent of counterfactual conditionals: “If you had taken any of that arsenic, you would have died” (see van Rooij (2006) for discussion).

a context as downward entailing involves a more sophisticated logical insight than identifying a particular argument as valid, since it requires the validation of an abstract pattern of argument. For example, identifying negation as a downward entailing context requires checking this schema, for arbitrary sentences “A” and “B”: If “A” entails “B” then “It is not the case that B” entails “It is not the case that A.” That is just the kind of abstract formal reasoning task on which humans perform worst. Contrast that with our high level of reliability in determining whether strings with negative polarity items are well-formed. Thus the evidence suggests that the unconscious logic in question is not at the service of the cognitive processes that normally produce conscious assent to sentences like (1). Such cases therefore fail to support a modification of the conclusions reached so far.

One special sort of case deserves separate discussion. Some metalinguistic sentences or thoughts look analytic for distinctive reasons. As observed in Chapter 2, even when a philosophical question is not itself metalinguistic, metalinguistic considerations can still help us to answer it.

Consider theoretical terms. We can understand the word “phlogiston” without believing phlogiston theory. Might we do so because we still believe that “phlogiston” is generally associated with that theory, just as one can understand a natural kind such as “gorilla” without believing the associated stereotype (“Gorillas are ferocious”) because one still believes that “gorilla” is generally associated with that stereotype (Putnam 1975)? However, such sociolinguistic beliefs are no more immune than logical beliefs from the challenge of theoretical unorthodoxy without change of meaning. If T is any version of phlogiston theory, someone can understand “phlogiston” and associate it with T without believing that it is generally associated with T, in the belief that “phlogiston” is and was generally associated not with T but with somewhat different versions of phlogiston theory. This is clear if T is a strong version of the theory. Even if T is a weak version, they may believe that the word is generally associated with a stronger version, and deny that it is *ipso facto* associated with T. On such grounds, they may even disbelieve that they themselves associate the word with T. Let such sociolinguistic beliefs be false; nevertheless, holding them is quite consistent with understanding “phlogiston.” It is futile to multiply disjuncts and restrictive clauses

in the hope of formulating a sociolinguistic claim so anodyne that anyone who understands “phlogiston” *must* accept it. The result will just be a complex theoretical claim that ordinary speakers can legitimately doubt, on the grounds that such matters are hard to determine.

A more minimalist line of argument for metalinguistic analyticities appeals to the connection between understanding and knowledge of reference. Suppose that someone understands this sentence:

(8) “Tree” applies to all and only trees.

Then they understand its constituent words, in particular “tree.” So they know what “tree” means. For common nouns, knowledge of meaning requires knowledge of application conditions. Consequently, they know that “tree” applies to all and only trees. Moreover, since knowledge entails belief, they also believe that “tree” applies to all and only trees. Thus, it seems, they should knowledgeably assent to (1). The argument generalizes to a large class of disquotational claims (the identity of the expression mentioned on the left-hand side with the one used on the right-hand side is crucial, since if they were distinct understanding of the latter would not entail knowledge about the former).

Nevertheless, those who understand (8) may refuse assent to it. Stephen is an example, since on his view a universally quantified biconditional with borderline cases for both sides is not definitely true. Indeed, some supervaluationists about vagueness even deny such disquotational principles for vague terms, such as “tree”. However erroneous such theories of vagueness, holding them is consistent with ordinary linguistic understanding of (8). If understanding really does involve tacit propositional knowledge of meaning, that knowledge may contradict conscious beliefs.

Let us grant for the sake of argument that understanding (8) entails knowing both that “tree” applies to all and only trees and that (8) means that “tree” applies to all and only trees. How then can one understand (8) without assenting to it? We lack direct conscious access to whatever tacit knowledge linguistic understanding is supposed to consist in, otherwise semantics as a branch of empirical linguistics would be much easier than it actually is. We consciously entertain the proposition that “tree” applies to all and only trees as

presented by sentence (8), or by the corresponding conscious thought “*tree*” *applies to all and only trees*. In tacitly knowing that “*tree*” applies to all and only trees (if we do), we may tacitly entertain that proposition under a quite different unconscious mode of presentation. Thus understanding-assent links fail for sentences of natural language and conscious thoughts:

- (UAl\*) Necessarily, whoever understands the sentence “‘Tree’ applies to all and only trees” assents to it.
- (UAt\*) Necessarily, whoever grasps the thought “*tree*” *applies to all and only trees* assents to it.

For if linguistic understanding involves tacit propositional knowledge of meaning, it presumably involves tacit assent to the relevant propositions under modes of presentation of some sort. Any tacit assent to the proposition that “*tree*” applies to all and only trees need not be to it under the modes of presentation that UAl\* and UAt\* require. The same difficulty arises even if we require only a disposition to assent, as in UDAI’ and UDAAt’.<sup>28</sup>

To determine in exactly what sense of “tacit knowledge,” if any, understanding does involve tacit propositional knowledge of meaning lies beyond the scope of this book. According to Gareth Evans (1985: 338–9):

Tacit knowledge of the syntactic and semantic rules of the language are [sic] not states of the same kinds as the states we identify in our ordinary use of the terms “belief” and “knowledge.” Possession of tacit knowledge is exclusively manifested in speaking and understanding a language; the information is not even potentially at the service of any other project of the agent, nor can it interact with any other beliefs of the agent (whether genuine beliefs or other tacit “beliefs”) to yield further beliefs. Such concepts as we use in specifying it are not concepts we need to suppose the subject to possess, for the state is inferentially insulated from the rest of the subject’s thoughts and beliefs.

Even if the contrast is less extreme than Evans argues, the lack of inferential integration is real, and crucial here. Of course, the

<sup>28</sup> See also Soames (1995) for relevant considerations.

ordinary notions of knowledge and belief may well provide appropriate templates for the construction of new notions of “tacit knowledge” and “tacit belief” of value to cognitive psychology. It can be theoretically rewarding to exploit the similarities between tacit knowledge and ordinary knowledge, but for present purposes it is the differences that matter.

Whatever the nature of tacit assent and dissent, no reflective intellectual discipline operates at the level of such assent and dissent, even if such a tacit level is necessary for its operation. Thus linguists’ tacit knowledge of their native language does not already satisfy the goal of linguistics. Similarly, philosophy as a discipline operates at the level of conscious reflection and public discussion, whatever their unconscious underpinnings. For present purposes, we may therefore restrict assent to conscious assent and maintain the generalization that there are no necessary links from understanding to assent, or even to dispositions to assent.

To summarize: The case for treating lack of a disposition to assent to (1) as lack of linguistic competence depends on the status of (1) as an elementary truth of deductive logic. But human deductive competence is far more sensitive than linguistic competence to high intelligence and advanced education. Deductive competence is a reflective skill, often painfully acquired and under one’s personal control. It is not insulated from one’s conscious theorizing. Thus deductive proficiency is not a precondition of linguistic competence. Links from linguistic understanding to assent or to dispositions to assent fail.

## 5

The argument of the last two sections was at the level of language, not thought. It was directed primarily against UAl' and UDAl', not UAt' and UDAt'. Could a theorist of thought maintain UAt' or UDAt' while acknowledging Peter and Stephen as counterexamples to UAl' and UDAl'?

For the sake of argument, thoughts are being individuated by a cognitive criterion fine enough to suit an epistemological conception of analyticity, so we may assume that when a speaker understands a sentence, they associate it with a unique thought, in the intimate way

in which we associate the sentence “Grass is green” with the thought *grass is green*. In particular, the speaker assents to the sentence if and only if they assent to the thought. Consider Stephen (the argument is parallel for Peter). Since Stephen understands “Every vixen is a vixen,” he associates it with a unique thought *t*. Thus Stephen assents to “Every vixen is a vixen” if and only if he assents to *t*. But Stephen is an acknowledged counterexample to UAI’; he does not assent to the sentence “Every vixen is a vixen.” Therefore he does not assent to *t*. Consequently, if *t* is the thought *every vixen is a vixen*, Stephen does not assent to the thought *every vixen is a vixen*, in which case he is also a counterexample to UAt’. Thus if Stephen is not a counterexample to UAt’, the thought he associates with the sentence “Every vixen is a vixen” is not the thought *every vixen is a vixen*.

There is a parallel argument for dispositions. Stephen is an acknowledged counterexample to UDAI’; he understands “Every vixen is a vixen” while having no disposition to assent to it. We may therefore assume that he is relevantly stable; thus in all relevant situations *t* is the unique thought he associates with the sentence. Thus Stephen has a disposition to assent to “Every vixen is a vixen” if and only if he has a disposition to assent to *t*. Therefore he has no disposition to assent to *t*. Consequently, if *t* is the thought *every vixen is a vixen*, Stephen has no disposition to assent to the thought *every vixen is a vixen*, in which case he is also a counterexample to UDAAt’. Thus if Stephen is not a counterexample to UDAAt’, the thought he associates with the sentence “Every vixen is a vixen” is not the thought *every vixen is a vixen*.

The upshot is that theorists of thought can maintain links from understanding to assent or dispositions to assent at the level of thought while abandoning them at the level of language only if they deny that the thought Peter or Stephen associates with the sentence “Every vixen is a vixen” is the thought *every vixen is a vixen*. They may either deny that Peter and Stephen grasp the thought *every vixen is a vixen* at all or assert that they grasp the thought by some means other than that sentence and assent to it, or at least have a disposition to assent.

The thought *every vixen is a vixen* is the thought we associate with (1). Thus the envisaged theorist of thought is claiming that the thought we associate with (1) differs from the thoughts Peter and Stephen associate with it, even though all of us understand (1) with

its usual meaning in English.<sup>29</sup> This need not imply that (1) is indexical, expressing different propositions in the contexts of different speakers, for thoughts are not being identified with propositions. You might use the sentence “He is hungry” (pointing at me), which you associate with a demonstrative thought *he is hungry* to express the very proposition I express using the sentence “I am hungry,” which I associate with the distinct thought *I am hungry*; you associate the sentence “I am hungry” with the same thought but use it to express a different proposition, that you are hungry. For all that has been said, Peter and Stephen use (1) to express the same proposition as we do. But on what basis are the thoughts Peter and Stephen associate with (1) being distinguished from the thought we associate with (1)?

One could simply use the word “thought” subject to the stipulation that the inferential differences between Peter, Stephen, and us *constitute* differences between the thoughts we associate with (1). But what is the point of such a stipulation? As seen above, the linguistic understanding of (1) we share with Peter and Stephen already suffices for them and us to articulate our disagreements in rational discourse; we are not merely talking past one another. In its small way, (1) determines a piece of the common intellectual heritage of mankind, something we share with Peter and Stephen in our very capacity to disagree over it. To insist that the thought we associate with (1) nevertheless differs from the thoughts Peter and Stephen associate with (1) is to undermine Frege’s requirement of the publicity of senses, and in particular thoughts.

If Peter and Stephen associate (1) with different thoughts from ours, should we not understand them better by translating their idiolects non-homophonically into ours? Presumably we should seek sentences other than (1) that we associate with the very thoughts they associate with (1), or at least sentences we associate with thoughts

<sup>29</sup> Neo-Fregeans such as Evans (1982: 40) sometimes claim that different speakers can achieve linguistic competence with the same proper name by associating it with different concepts (modes of presentation) of the same object. On the view envisaged in the text, phrases such as “the thought *every vixen is a vixen*” or “the concept *every*” presumably are indexical, since they refer to the thought or concept that the speaker associates with the italicized expression. Discussions of concept possession tend to use such phrases freely, without attention to such indexicality. On the envisaged view, they may require consequent revision.

more similar to the thoughts they associate with (1) than is the thought we associate with (1), and translate the dissent from (1) in their mouths as dissent from those other sentences in our mouths. But the use of such a translation scheme would be intellectually disreputable, just because it would involve a refusal to acknowledge the full challenge that Peter and Stephen have issued to (1) in our mouths, not just in theirs. However mistaken their challenge, it is real. They are quite explicit that they are challenging the thought we associate with (1), and that we should apply no non-homophonic translation scheme when interpreting their dissent from (1). To insist on applying such a non-homophonic translation scheme to them in the teeth of their protests would be to treat them less than fully seriously as human beings, like patients in need of old-fashioned psychiatric treatment, whose words are merely symptoms. The claim that Peter and Stephen associate (1) with different thoughts from ours repackages our disagreement with them in a way that makes it sound less threatening than it really is. It misleadingly bundles together logical and semantic differences, without any genuine unification of the two categories. To call the logical disagreement a difference in associated “thoughts” is an advertising trick. Since a homophonic reading of (1) in the mouths of Peter and Stephen is more faithful to their intentions than is any non-homophonic reading, they associate (1) with the same thought as we do in any relevant sense of “thought.”

Naturally, when Peter dissents from “Every F is a G,” we may decide in the light of his logical unorthodoxy to store only the information that either not every F is a G or there are no Fs. But this is not a non-homophonic *translation*, any more than it is when someone notorious for exaggeration says “At least six thousand people went on the march” and we decide to store only the information that at least one thousand people went on the march. By “six thousand” the speaker did not mean what we mean by “one thousand.” If exactly one thousand people went on the march he spoke falsely, not truly, for he was speaking English. Since we do not fully trust him, when he asserted one thing we stored only something weaker. Similarly, since we do not fully trust Peter, we do not store exactly what he asserts. If there were no Fs, he spoke falsely, not truly, for he was speaking English. Our lack of trust in Peter and Stephen’s logic skills is quite consistent with reading their utterances homophonically.

Peter and Stephen are counterexamples to UAt' and UDAt'. The links from understanding to assent, or even to dispositions to assent, fail for thought as they do for language.

## 6

How do the considerations of preceding sections apply to traditional paradigms of analyticity? Consider:

(9) Every vixen is a female fox.

Given that “vixen” is synonymous with “female fox,” (9) results from substituting synonyms for synonyms in the logical truth (9). Hence (9) is synonymous with (1): it is Frege-analytic but not itself a logical truth. We can expect the arguments of previous sections against links from understanding to assent or dispositions to assent for examples like (1) to work at least as strongly for examples like (9). Let us check this.

We may try to reduce discussion of (9) to discussion of (1), on the grounds that the concept *vixen* just is the concept *female fox*. Thus the thought *every vixen is a female fox* just is the thought *every vixen is a vixen* (since thoughts are composed of concepts). To grasp, assent to or know a thought is just to have a relation to that thought. Consequently, to grasp, assent, or know *every vixen is a female fox* just is to grasp, assent, or know *every vixen is a vixen*. At the level of thought, the previous discussion carries over automatically. For example, in being counterexamples to the understanding-assent link for the thought *every vixen is a vixen*, Peter and Stephen are *ipso facto* counterexamples to the understanding link for the thought *every vixen is a female fox*.

At the level of language, the reduction is slightly more complicated: “vixen” and “female fox” are distinct expressions even if they are associated with the same concept. Someone can understand “female fox” without understanding “vixen.” Conversely, someone can understand “vixen” without understanding “female fox”: for instance, a native speaker of another language who is learning English understands “vixen,” because she was taught it as a synonym for a word in her native language, but has not yet encountered “female” and

“fox.” If she has mastered the construction “Every . . . is a *–*,” she can understand (1) without being in a position to understand (9). Someone who understands neither (1) nor (9) can assent to one of them without assenting to the other, on the testimony of someone else who tells him that the former is true without telling him that the latter is true. Nevertheless, we might try arguing that whoever understands (9) will take just the same attitudes to it as to (1).

The argument is this. Suppose that someone understands (9) (as always, with its normal English meaning). Thus she associates it with the thought *every vixen is a female fox*. Consequently, she takes an attitude Al (such as assent or knowledge) to (9) if and only if she takes the corresponding attitude At to the thought *every vixen is a female fox* at the level of thought (in preceding sections, Al and At were equated). Our speaker also understands (1), because it is composed entirely out of words (“vixen”) and modes of construction (“every . . . is a *–*”) which she understands in understanding (9). Thus she associates (1) with the thought *every vixen is a vixen*. Consequently, she takes Al to (9) if and only if she takes At to the thought *every vixen is a vixen*. For the reason already given, the thought *every vixen is a vixen* is the thought *every vixen is a female fox*. Therefore she takes At to the thought *every vixen is a vixen* if and only if she takes At to the thought *every vixen is a female fox*. It follows that she takes Al to (9) if and only if she takes Al to (1). Thus, with respect to speakers who understand (9), discussion of (9) reduces to discussion of (1).

Whether or not the concept *vixen* is the concept *female fox*, the reduction succeeds for Peter and Stephen, since they use the concepts interchangeably and do understand (9). They are counterexamples to epistemological analyticity for (9) just as much as they are for (1), at the levels of both thought and language.

The assumption that the concept *vixen* is the concept *female fox* is controversial. Burge (1978) has built on a point of Mates (1952) to argue that synonyms cannot always be substituted for synonyms *salva veritate* in belief ascriptions. Thus someone under the misapprehension that the term “vixen” also applies to immature male foxes may believe that every vixen is a vixen without believing that every vixen is a female fox. Burge argues powerfully against attempts to reconstrue such beliefs as metalinguistic. Does this speaker assent to the thought *every vixen is a vixen* without assenting to the thought *every vixen is a female fox*? If so, the thoughts are distinct (which is

compatible with the identity of the proposition that every vixen is a vixen with the proposition that every vixen is a female fox), and the concept *vixen* is not the concept *female fox*.

To make a case more like those of Peter and Stephen, we can imagine that our speaker is quite familiar with the dictionary definition of “vixen” as “female fox.” He also knows that dictionaries give a second definition of “vixen” as “quarrelsome woman.” However, unlike most of us, he does not believe that these are two senses of “vixen.” Rather, he thinks that “vixen” in its primary sense applies to both female foxes and quarrelsome women. He may defend his view with sophisticated arguments from the philosophy of language, although this is not essential. He denies (9), intending “vixen” in the public sense in which it applies at least to female foxes.

Our imaginary speaker is not so different from actual natives speakers of English who deny that a man who has lived with a partner for several years without getting married is a bachelor, or assert that someone who underwent a sex-change operation after giving birth is a mother without being a female parent.<sup>30</sup> Suppose that they are in fact mistaken; “bachelor” has the same intension as “unmarried man” and “mother” has the same intension as “female parent.” Thus they are mistaken about the meaning of the English words “bachelor” and “unmarried.” Nevertheless, they fall well within the range of permissible variation for linguistically competent speakers. They are only giving more weight than others to an inclination that most speakers feel in some degree to classify the cases that way. Without regarding them as having spoken parrot-fashion, we report their beliefs using the words “bachelor” and “unmarried.” We classify them as believing that some unmarried men are not bachelors and that some mothers are not female parents because we interpret them as having used the words with their normal English meanings, despite their errors. That is how they intend to be interpreted, not as using the words with idiosyncratic senses.<sup>31</sup> If we believe that all unmarried

<sup>30</sup> Compare Harman (1999: 151) on problems in analyzing “bachelor” as “unmarried adult male” and Nozick (2001: 135–6) on the non-synonymy of “mother” and “female parent.”

<sup>31</sup> One problem with interpreting speakers as all speaking their own idiolects is that it tends to undermine testimonial knowledge: if Y gets some knowledge from X and passes it on to Z in the same words, they do not mean in Y’s mouth what they meant in X’s.

men are bachelors and all mothers are female parents, we therefore classify their beliefs in question as untrue, for the belief that some unmarried men are not bachelors is true if and only if some unmarried men are not bachelors, and the belief that some mothers are not female parents is true if and only if some mothers are not female parents. Given that we correctly interpret them as using the words with their normal English meanings, they understand the words in the relevant sense of “understand.” Although they are ignorant of some facts about the normal English meanings of the words, such ignorance is quite compatible with linguistic competence (which is why native speakers of English take university courses in the semantics of English). Arguably, their error is not primarily semantic: they have the semantic belief that the word “bachelor” does not apply to all unmarried men because they have the non-semantic belief that some unmarried men are not bachelors and the semantic knowledge that “bachelor” applies only to bachelors; they have the semantic belief that the word “mother” does not apply only to female parents because they have the non-semantic belief that some mothers are not female parents and the semantic knowledge that the word “mother” applies to all mothers.

Such cases also help answer the objection to examples such as those in this chapter that the awkward subject who consciously denies that P also has unconscious, semantically derived knowledge (or belief) that P. When a competent native speaker denies that every unmarried man is a bachelor, the postulation of unconscious knowledge (or belief) that every unmarried man is a bachelor serves no good explanatory purpose. The speaker tends to apply “bachelor” to something once they have applied “unmarried” and “man” to it, but the tendency is defeasible. Such defeasible connections can be explained without postulation of unconscious belief in a universal generalization. In such cases, there need be no hint of the cognitive dissonance or tension that one might expect from a direct contradiction between conscious and unconscious beliefs. Given that there is no contradicted unconscious knowledge in these simple cases, it is not clear what better reason there is supposed to be in postulating it for more complex cases either.

Suppose, given the considerations above, that the concept *vixen* is not the concept *female fox*. Then the claim of epistemological analyticity is even worse off for (9) than it is for (1), at the levels of both

thought and language. Logically orthodox subjects can understand (9) and grasp the thought *every vixen is a female fox* while refusing to assent. In that case, they will also reject the corresponding inference rule with instances of the form “*a* is a vixen; therefore *a* is a female fox” (and conversely); likewise at the level of thought.<sup>32</sup>

The underlying style of argument against links from understanding to assent or dispositions to assent is quite general. For each candidate one must still find appropriate counterexamples: since they are most convincing when unorthodoxy on the point at issue is amply compensated by orthodoxy on related points, no one counterexample will suit all cases. Nevertheless, with a little ingenuity one always succeeds.<sup>33</sup>

<sup>32</sup> Peter and Stephen assent to the conclusion of this inference rule whenever they assent to its premise. For some subtler problems it raises for them see Williamson (2006b: 33–4).

<sup>33</sup> Another application of the present style of reasoning is to claims that sorites paradoxes reveal incoherence in vague concepts. Thus Dummett (1975a) argues that observational predicates in natural language are governed by rules that infect the language with inconsistency: for example, to understand “looks red” one must be willing to apply a tolerance principle by which one can infer from “*x* is visually indiscriminable from *y*” and “*x* looks red” to “*y* looks red,” which generates sorites paradoxes because visual indiscriminability is non-transitive. More recently, Roy Sorensen (2001) has argued that linguistic competence with vague terms involves willingness to make inferences such as that from “*n* seconds after noon is noonish” to “*n* + 1 seconds after noon is noonish,” which commits us to inconsistent conclusions by sorites reasoning (given our other commitments, such as “Noon is noonish” and “Midnight is not noonish”). Matti Eklund (2002) defends a similar account of both sorites and semantic paradoxes. There are no such requirements on linguistic competence and concept possession. An ordinary speaker of English who understands “looks red” and “noonish” and has the concepts *looks red* and *noonish* in the normal way but then rejects the relevant tolerance principles in the light of the sorites paradoxes does not thereby cease to understand those expressions or to have those concepts. She might treat the premises of the tolerance principles as providing good defeasible evidence for their conclusions, without even being *disposed* to expect long chains of such reasoning to preserve truth; this attitude seems to be less than Dummett, Sorensen, and Eklund require for competence, since it is insufficient to render sorites paradoxes puzzling. In any case, even if a whole community of speakers is disposed to treat tolerance principles as obviously fallacious, it can still have terms like “looks red” and “noonish” that are just as vague as ours; speakers’ acceptance of tolerance principles is quite inessential to vagueness.

In principle, we could also explore putative links from understanding of one sentence to (dispositions to) assent to another sentence or a thought, or from grasp of one thought to (dispositions to) assent to another thought. In practice, such candidates fall to objections very similar to those already raised. Details are therefore omitted.

## 7

Old theories tend to survive refutation in the absence of new theories to take their place. Despite all the evidence against the existence of links from understanding to assent or dispositions to assent, it can be hard to resist the idea that there *must* be such links, otherwise the distinction between understanding and not understanding would dissolve: speakers who all understood the same term might have nothing substantive in common to constitute its shared meaning. For example, in the case of moral vocabulary, which he treats as representative, Frank Jackson (1998: 132) writes:

Genuine moral disagreement, as opposed to mere talking past one another, requires a background of shared moral opinion to fix a common, or near enough common, set of meanings for our moral terms. We can think of the rather general principles that we share as the commonplaces or platitudes or constitutive principles that make up the core we need to share in order to count as speaking a common moral language.<sup>34</sup>

<sup>34</sup> Jackson's application of the Ramsey-Carnap-Lewis method for defining theoretical terms to moral vocabulary (and more generally in his program of conceptual analysis) requires not merely some agreed role for moral terms but an agreed role specific enough to be uniquely instantiated: this further assumption is criticized at Williamson (2001: 629–30). Jackson's reply on this point (2001: 656) reiterates something like the assumption in the quoted passage. He also misunderstands the objection by falsely supposing that the claim that we can mean the same by a word and disagree radically about its application restricts the disagreement to what occupies the roles, rather than the roles themselves, however one imagines the latter as demarcated. For criticism of the application of the Ramsey-Carnap-Lewis method in Boghossian (2003) see Williamson (2003a). In general, if the platitudes are weak, as we have every reason to expect, many different candidates will satisfy them. Call these the *admissible candidates*. For simplicity, think of them as properties (more accurately, they are  $n$ -place sequences of properties and relations, where  $n$  is the number

Jackson's only argument for these claims is failure to see an alternative.

The notion of a shared language is vague (Jackson does not suggest otherwise). There can be sorites series of speakers in which each seems to be speaking the same language as the next but the first is clearly not speaking the same language as the last.<sup>35</sup> One reaction is that there is no such thing as a shared language, a conclusion endorsed in some form by both Noam Chomsky and Donald Davidson. Similarly, Margaret Thatcher once claimed "There is no such thing as society," and one can certainly construct sorites series in her support. But almost everything looks vulnerable to sorites series; they are a poor way to establish non-existence. Whatever exactly shared languages are, they are no mere illusion. We can follow Jackson in asking how they are possible. But there is an alternative to his answer.

---

of primitive predicates to which the method is being applied). The conjunction or disjunction of these admissible candidates will often not itself be an admissible candidate. Schematic example: let the platitudes be "All Fs are electrons," "Some electrons are Fs" and "Some electrons are not Fs," where the method is being applied to "F"; the conjunction of the admissible candidates is the empty property, which does not satisfy the second platitude and so is inadmissible; their disjunction is the property of being an electron, which does not satisfy the third platitude and so is inadmissible. The non-uniqueness problem for the Ramsey-Lewis-Carnap *definiens*, in effect "the property that satisfies the platitudes," is *not* that it is vague which property it denotes but that it definitely fails to denote any property at all, since many properties definitely satisfy the platitudes; neither supervaluationism nor any other theory of vagueness rescues the definition. A modified description such as "the most natural property that satisfies the platitudes" may still not solve the problem – perhaps several admissible candidates are equally natural and more natural than any others, or for every admissible candidate there is a more natural one – and in any case raises the question why the Ramsey-Carnap-Lewis method is being applied to some terms but not to the highly theoretical term "natural" itself (otherwise the problem simply recurs for "natural"). It is a mistake to assume that such problems are really problems for the linguistic practice itself rather than for the appeal to platitudes, for that is to assume that the platitudes exhaust what the practice does to secure reference for the predicate. Uses of the predicate to make controversial claims may also play a role in determining its reference, although not a naïvely descriptivist role (the account in Chapter 8 will permit this). The method of platitudes rashly throws such information away.

<sup>35</sup> Williamson (1990: 137–41) discusses sorites series for languages.

What binds together uses of a word by different agents or at different times into a common practice of using that word with a given meaning? This is an instance of a more general type of question: what binds together different events into the history of a single complex object, whether it be a stone, a tree, a table, a person, a society, a tradition, or a word? In brief, what makes a unity out of diversity? Rarely is the answer to such questions the mutual similarity of the constituents. Almost never is it some invariant feature, shared by all the constituents and somehow prior to the complex whole itself – an indivisible soul or bare particular. Rather, it is the complex interrelations of the constituents, above all, their causal interrelations. Although we should not expect a precise non-circular statement of necessary and sufficient conditions for the unity in terms of those complex interrelations, we have at least a rough idea of what it takes. The similarity of the constituents is neither necessary nor sufficient; different constituents can play different but complementary roles in constituting the unity: both events in the head and events in the heart help constitute the life of a person. The idea that a shared understanding of a word requires a shared stock of platitudes depends on the assumption that uses of a word by different agents or at different times can be bound together into a common practice of using that word with a given meaning only by an invariant core of beliefs. But that assumption amounts to one of the crudest and least plausible answers to the question of what makes a unity out of diversity. In effect, it assumes that what animates a word is a soul of doctrine.<sup>36</sup>

As Kripke and Putnam argued, different speakers can make asymmetric contributions to binding together different uses of a word into a common practice of using it with a given meaning. The paradigm is their description of the role of scientific experts in fixing the reference of natural kind terms. Even if they oversimplified the relation between natural kind terms in natural language and scientific theory, a more refined account will still respect the division of linguistic labor, for distinctions between levels of expertise are observable even within the pre-scientific use of natural kind terms. Contrary to some of Putnam's less careful formulations, no canonical list of "criteria"

<sup>36</sup> A similar point is made in Schroeter and Schroeter (2006). More generally, the research program that these authors are pursuing has points in contact with the ideas of the present chapter.

for the application of the term need be available even to the most expert members of the community. Speakers may simply differ from each other in various ways in their ability to distinguish between members and non-members of the relevant kind.

The underlying insight is relevant far beyond the class of natural kind terms, as Burge observed. Even where we cannot sensibly divide the linguistic community into experts and non-experts, the picture of a natural language as a cluster of causally interrelated but constitutively independent idiolects is still wrong, because it ignores the way in which individual speakers defer to the linguistic community as a whole. They use a word as a word of a public language, allowing its reference in their mouths to be fixed by its use over the whole community.<sup>37</sup> No asymmetries in sociolinguistic status between individual speakers are required. For instance, if I classify a shade close to orange as “red” but subsequently discover that it is classified as “not red” by most native speakers of English whose eyesight is as good as mine, I may rationally admit that I was wrong without conceding that either I misunderstood the word “red” or my visual system was abnormal or malfunctioning. One can know that “red” means *red* without being infallible as to exactly which shades count as shades of red. Even if I obstinately insist that I am right and the rest are wrong in this particular case, my assumption that “red” in my mouth is inconsistent with “not red” in theirs shows that I intend my use of “red” to be treated as the use of a word of a public language. That its reference is fixed by the pattern of use over the whole community does not entail that the majority must be right in any given case: reference can supervene on underlying facts in ways far from transparent to native speakers.

The unity of a linguistic practice, like the unity of other complex objects, has both synchronic and diachronic aspects. As usual, causal continuity is necessary but not sufficient for diachronic unity. Anaphoric pronouns constitute one paradigm of such unity: the reference of later tokens is parasitic on the reference of earlier tokens; the identity of reference results from collusion, not coincidence. Over a longer timescale, the historical chains that preserve the reference of names represent a similar form of diachronic unity. Written testi-

<sup>37</sup> If the term is indexical, what is fixed by use over the whole community is not the content but the character in the sense of Kaplan (1989). For the bearing of this on communication in a vague language see Williamson (1999b: 512–14).

mony and verbal testimony preserved in memory depend on such reference-preserving links. As usual, the intention to preserve reference is not guaranteed to succeed, but success is the default (Kripke 1980).

Such diachronic links can hold non-trivially even for the linguistic or conceptual practice of an isolated individual. Contrary to some readings of Wittgenstein's private language argument, what seems right to the isolated individual need not be right, given their overall use dispositions: even at the individual level, reference can supervene on underlying facts in ways far from transparent to the subject. The point of the social determination of meaning is not that meaning can never be determined individually, but that, when an individual does use a shared language as such, individual meaning is parasitic on social meaning.

A complex web of interactions and dependences can hold a linguistic or conceptual practice together even in the absence of a common creed that all participants at all times are required to endorse. This more tolerant form of unity arguably serves our purposes better than would the use of platitudes as entrance examinations for linguistic practices.

Evidently, much of the practical value of a language consists in its capacity to facilitate communication between agents in epistemically asymmetric positions, when the speaker or writer knows about things about which the hearer or reader is ignorant, perhaps mistaken. Although disagreement is naturally easier to negotiate and usually more fruitful against a background of extensive agreement, it does not follow that any particular agreement is needed for disagreement to be expressed in given words. A practical constraint on useful communication should not be confused with a necessary condition for literal understanding. Moreover, the practical constraint is holistic; agreement on any given point can be traded for agreement on others. The same applies to principles of charity as putatively constitutive conditions on correct interpretation: imputed disagreement on any given point can be compensated for by imputed agreement on others.<sup>38</sup>

<sup>38</sup> Davidson famously endorses a holistic principle of charity while rejecting the analytic-synthetic distinction (2001: 144–9). See Chapter 8 for more discussion of charity. Of course, he takes the notion of a shared language less seriously than here (Davidson 1986).

It is far easier and more rewarding to discuss the existence of true contradictions with a dialetheist such as Graham Priest than creationism with a Christian fundamentalist or Holocaust denial with a neo-Nazi.<sup>39</sup> The difficulty of engaging in fruitful debate with fundamentalists or neo-Nazis is not plausibly attributed to some failure of linguistic understanding on their part (or ours); it arises from their willful disrespect for the evidence. Such difficulty as there is in engaging in fruitful debate with dialetheists provides no significant reason to attribute to them (or us) a failure of linguistic understanding. Competence with the English language no more requires acceptance of some law of non-contradiction or any other logical law than it requires acceptance of the theory of evolution or the historical reality of the Holocaust.

We cannot anticipate all our disagreements in advance. What strike us today as the best candidates for analytic or conceptual truth some innovative thinker may call into question tomorrow for intelligible reasons. Even when we hold fast to our original belief, we can usually find ways of engaging rationally with the doubter. If a language imposes conditions of understanding that exclude such a doubt in advance, as it were in ignorance of its grounds, it needlessly limits its speakers' capacity to articulate and benefit from critical reflection on their ways of thinking. Such conditions are dysfunctional, and natural languages do not impose them.<sup>40</sup> Similarly, conceptual practices do better not to restrict in advance their capacity for innovation.

There is, of course, a distinction between understanding a word and not understanding it. One can lack understanding of a word through lack of causal interaction with the social practice of using that word, or through interaction too superficial to permit sufficiently fluent engagement in the practice. But sufficiently fluent engagement in the practice can take many forms, which have no single core of agreement.<sup>41</sup>

<sup>39</sup> For examples of rational debate for and against a law of non-contradiction see Priest, Beall, and Armour-Garb (2004).

<sup>40</sup> W. B. Gallie's intriguing account of the positive function of "essentially contested concepts" is relevant here; his examples are "the concepts of a religion, of art, of science, of democracy and of social justice" (1964: 168).

<sup>41</sup> Someone who understands a word without being disposed to utter it (perhaps because they find it obscene or unpronounceable) can still count as sufficiently

If we picture speaking the same language in this way, how should we picture meaning the same thing? There is no quick generalization from the former to the latter. Different uses of the same word must be causally related, at least indirectly.<sup>42</sup> Creatures who are causally unrelated to us cannot use our word “not”; at best they can use a word exactly like our word in its general syntactic, semantic, and phonetic properties. But, on the usual view, their word can in principle be synonymous with ours. Synonymy does not entail causal relatedness.

Expressions are synonymous when they have exactly the same semantic properties. Fortunately, the tradition of truth-conditional semantics provides us with a rich store of such properties, if we take it seriously as a branch of linguistics and put aside Quinean reservations.

Two paradigms of a semantic property are the extension of a predicate, the set of things to which it applies, and its intension, the function that takes each circumstance of evaluation (say, an ordered pair of a world and a time) to the extension of the predicate with respect to that circumstance. For the purposes of compositional semantics, this approach can be generalized to expressions of other grammatical categories, so that they have intensions too. Thus synonymy entails at least sameness of intension. That is still a rather coarse-grained criterion, since it does not reflect internal compositional structure: “5 + 7” and “9 + 3” have the same intension. We can go more fine-grained by associating expressions with trees whose nodes correspond to their semantically significant constituents, each node being decorated with the content of the corresponding constituent; the branching structure of the tree encodes the constituency structure of the expression. Thus synonymy entails at least sameness of associated tree. This criterion is similar to Carnap’s notion of intensional isomorphism (1947: 56). In this sense not even “vixen” and “female fox” are synonymous, since they differ in semantically significant structure, unless the account can be applied at a level of deep logical form at which they turn out to have the same constituents. Something like intensional isomorphism can serve as a criterion for sameness of content expressed in a given context of utterance.

---

engaged in the practice of using it. The account should also be read so as to allow for understanding of dead languages.

<sup>42</sup> On the metaphysics of words see Kaplan (1990).

An expression brings its linguistic meaning to a context rather than having that meaning made up in the context. Thus “I” as used by TW does not have the same linguistic meaning as “TW,” even though they have the same content (since they are unstructured rigid designators of the same object). Rather, “I” as used by TW is identical in linguistic meaning with “I” as used by any other competent speaker of English. Thus a better approximation to the linguistic meaning of an expression is its character in the sense of Kaplan (1989), the function taking each context of utterance to the content of the expression in that context.

We might go still further. For instance, so far “and” and “but” come out synonymous, since they are simple expressions that make the same contribution to truth-conditions. We might distinguish their meanings by adding as further semantic properties conventional implicatures, themselves individuated like characters.

Even without conventional implicatures, once content is individuated by intensional isomorphism, the conception of linguistic meaning as character is already exquisitely fine-grained. Nevertheless, if semantic theory discovers a need to attribute still more semantic properties, or to revise the framework already sketched, sameness with respect to the newly identified semantic properties will be required for synonymy. In any case, we need not try to circumscribe in advance exactly what properties semantic theory will need to recognize.

The point is methodological. Whether an expression in one language is synonymous with an expression in another language is not a matter of whether the two speech communities associate similar beliefs with the expressions. Rather, the practices of each community (including their beliefs) determine the semantic properties of its expressions. Synonymy is the identity of the properties so determined, irrespective of similarities in belief. It is consistent with large differences in belief (just as very different distributions can have the same mean), and non-synonymy is consistent with much smaller differences in belief (just as very similar distributions can have different means). In particular, synonymy is consistent with the total absence of shared platitudes.

The synonymy of two expressions does not entail that competent speakers treat them interchangeably, as noted in chapter 3. Someone can understand “furze” and “gorse” by learning them from ostension of different samples without appreciating their synonymy. In some

cases, even competent speakers who know two expressions to be synonymous will not treat them interchangeably. For example, the slang word “gob” means the same as “mouth,” but competent speakers are normally sensitive to whether the social context makes “gob” (but not “mouth”) inappropriate. Such differences in register are linguistic but not semantic. Consequently, knowing the meaning of an expression does not automatically qualify one for full participation in the practice of using it. Someone who acquires the word “gob” just by being reliably told that it is synonymous with “mouth” knows what “gob” means without being fully competent to use it. One does not achieve full competence with a sentence of a foreign language by learning its meaning from a phrasebook without knowing which constituent contributes what to that meaning. For a less obvious case, consider empty terms. Arguably, “phlogiston” fails to refer with respect to any circumstance of evaluation (since it designates rigidly, if at all) and any context of utterance (since it is non-indexical); it is semantically atomic and has no conventional implicatures. Those facts may completely determine its semantics, strictly speaking. Nevertheless, knowing them alone does not qualify one to participate in the linguistic practice of using “phlogiston,” since they do not distinguish it from empty terms associated with other failed theories. Although no particular piece of knowledge is necessary for participation, such abject ignorance is not sufficient. We should resist the temptation to build all qualifications for participation in the practice of using a term into its meaning, on pain of turning semantic theory into a ragbag of miscellaneous considerations (even the inclusion of conventional implicature is marginal).

What of concepts? Presumably, thinkers causally unrelated to us could have the concept *not*. Hence sameness of concept does not entail causal relatedness; it is closer to sameness of meaning than it is to sameness of word. If so, the concept *furze* may well just be the concept *gorse*. If thoughts are composed of concepts in the obvious way, then the thought *all furze is gorse* just is the thought *all furze is furze*, and whoever assents to the latter *ipso facto* assents to the former. We may sometimes be unable to determine whether we are employing two concepts or one. That makes the individuation of thoughts and concepts less accessible to the thinker than many theorists of thought have wished. For the sake of greater (but still imperfect) accessibility, they might therefore switch to individuating

concepts more like words than like meanings. In any case, the argument against epistemological analyticity at the level of thought has already been explained, in Section 5.

## 8

At this point, a friend of epistemological analyticity may suspect that the mistake was to go for the idea that understanding is somehow *psychologically* sufficient for assent. Instead, the suggestion is, we should go for the idea that understanding is somehow *epistemologically* sufficient for assent.<sup>43</sup> Externally, Peter and Stephen are in a position to know (or to assent with justification). They seem to be willfully and perversely turning their backs on knowledge that is available to them. It is there for the taking, but they are psychologically blocked from taking it.

We must be careful about the source of the blockage. Suppose that it is lack of logical insight. Although Peter and Stephen grasp the thought *every vixen is a vixen*, they lack the logical insight to know *every vixen is a vixen*. Other people just like Peter and Stephen except for having more logical insight do know *every vixen is a vixen*. Anyone who grasps the thought *every vixen is a vixen* and has a modicum of logical insight can know *every vixen is a vixen*. That story assigns no special role to grasp of concepts, beyond the usual role that grasping any thought plays as a precondition for knowing it: the decisive role is assigned to logical competence, not conceptual competence. For conceptual competence to play the decisive role, something like this is needed:

(KUt') Whoever knows *every vixen is a vixen* in the normal way does so simply on the basis of their grasp of the thought.

(Understand “on the basis of” more like “by an exercise of” than like “by inference from.”) Similarly, for semantic competence to play the decisive role, something like this is needed:

<sup>43</sup> Some rationalist defenders of intuition seem to have something like this in mind.

(KUI') Whoever knows “Every vixen is a vixen” in the normal way does so simply on the basis of their understanding of the sentence.

KU $t'$  and KU $l'$  may be plausible at first sight. They do not imply that whoever understands the sentence or grasps the thought has a disposition to assent to it, let alone to know it.

What do the definite descriptions “their grasp of the thought” in KU $t'$  and “their understanding of the sentence” in KU $l'$  denote? There are thick and thin candidates. The thin candidates are the mere fact that they grasp the thought and the mere fact that they understand the sentence respectively. The thick candidates are the underlying facts that constitute the respective thin candidates, the facts that realize this particular subject’s understanding at this particular time. The thin candidates are exactly similar for any two people who grasp the thought or understand the sentence, since they have the same property of grasping the thought or understanding the sentence. The thick candidates may differ between any two people who grasp the thought or understand the sentence, since different underlying facts can constitute their doing so. These characterizations are schematic, but will do for present purposes.

Suppose that the definite descriptions in KU $t'$  and KU $l'$  denote the thick candidates. KU $t'$  and KU $l'$  remain somewhat plausible on this reading. Then, given the holistic picture of concept possession and linguistic understanding in previous sections, KU $t'$  and KU $l'$  have much less epistemological significance than might have been hoped. The facts that constitute your understanding of a given sentence include various cognitive capacities that are not in general necessary for understanding that sentence, but help to make up your particular competence with it. For example, the facts that constitute Peter’s understanding of (1) include his logical capacities; the facts that constitute Stephen’s understanding of (1) include his rather different logical capacities. The bases cited in KU $l'$  and KU $t'$  include cognitive capacities that are not in general necessary for understanding the sentence or grasping the thought. Thus the thick candidates are too thick to yield bases for analyticity; they involve cognitive capacities that are not semantic or conceptual in any relevant sense.

Suppose instead that the definite descriptions in KU $t'$  and KU $l'$  denote the thin candidates. But they are not the bases in any useful

sense for knowing *every vixen is a vixen* or “Every vixen is a vixen” in the normal way, although confusion with the thick candidates may suggest otherwise. The thin candidates imply no specific logical capacity at all, as Peter, Stephen and others show. It is not as though in such cases the subject’s understanding quietly tells them to assent but they override the advice; it is providing no such advice to be overridden. For the imagined overridden advice is a metaphor for the hypothesis of overridden dispositions to assent, dispositions necessary for understanding; that hypothesis was rejected in Section 4. By itself, thin understanding cannot guide our assent. Consequently, understanding in the thin sense provides no basis for assent to anyone. Of course, understanding is a precondition for knowing, and in that sense may be *part* of the basis for knowing, but that point is quite general; it is neutral between the analytic and the synthetic. Although the combination of understanding in the thin sense with the right bit of elementary but not universal logical competence is a basis for knowing (1), that point neither explains why logical knowledge is available in the armchair nor makes it distinctively conceptual or semantic. By themselves, the thin candidates are too thin to be bases for knowledge.

We could try eliminating the talk of bases, for instance in formulations like these:

- (AJt') Whoever grasps the thought *every vixen is a vixen* and assents to it does so with justification.
- (AJl') Whoever understands the sentence “Every vixen is a vixen” and assents to it does so with justification.<sup>44</sup>

But such principles are false, since someone who assents because his father told him not to does so without even defeasible justification. The obvious way to avoid such counterexamples and make the connection with conceptual or semantic competence is to qualify “assents to it” by “on the basis of that grasp [understanding].” But that returns us to the difficulties of KU<sup>t</sup> and KU<sup>l</sup>.

<sup>44</sup> The intended differences between assenting with justification in AJt' and AJl' and being justified in assenting in UJt and UJl are that (i) the former but not the latter entails assent and (ii) the assent in the former must be appropriately sensitive to the justification.

The problem is general. The idea that, in the cases at issue, understanding is epistemologically sufficient for assent is the idea that assent on the basis of understanding has the desired positive epistemic status. But once we disambiguate “understanding” between thick and thin candidates, we can see that the thin candidates are too thin to be bases for assent while the thick candidates are not purely semantic or conceptual. The attempt to base the epistemology of obvious truths such as (1) and (9) on preconditions for understanding them rests on a false conception of understanding.

Linguistic competence plays the same role when we know “Vixens are female foxes” as when we know “There is a vixen in the garden.” It does not gain a role just because perception loses one. The contribution of linguistic competence amounts to this: you won’t get very far if you conduct your inquiry in a language you don’t understand. Of course, that goes for *any* inquiry.

The following chapters develop a quite different account of the nature of at least some philosophical knowledge, on which linguistic and conceptual competence play only this background role, and philosophical beliefs are much less distinctive in nature than many philosophers like to think. We start with knowledge of metaphysical possibility and necessity.

# 5

## Knowledge of Metaphysical Modality

---

### 1

Philosophers characteristically ask not just whether things are some way but whether they could have been otherwise. What could have been otherwise is *metaphysically contingent*; what could not is *metaphysically necessary*. We have some knowledge of such matters. We know that Henry VIII could have had more than six wives, but that three plus three could not have been more than six. So there should be an epistemology of metaphysical modality.

The differences between metaphysical necessity, contingency, and impossibility are not mind-dependent, in any useful sense of that frustrating phrase. Thus they are not differences in actual or potential psychological, social, linguistic, or even epistemic status (Kripke (1980) made the crucial distinctions). One shortcut to this conclusion uses the plausible idea that mathematical truth is mind-independent. Since mathematics is not contingent, the difference between truth and falsity in mathematics is also the difference between necessity and impossibility; consequently, the difference between necessity and impossibility is mind-independent. The difference between contingency and non-contingency is equally mind-independent; for if  $C$  is a mind-independently true or false mathematical conjecture, then one of  $C$  and its negation conjoined with the proposition that Henry VIII had six wives forms a contingently true conjunction while the other forms an impossible conjunction, but which is which is mind-independent. To emphasize the point, think of the mind-independently truth-valued conjecture as evidence-transcendent, absolutely undecidable, neither provable nor refutable by any means. Thus the epistemology of metaphysical modality is one of mind-independent truths.

Nevertheless, doubts begin to arise. Although philosophers attribute metaphysical necessity to mathematical theorems, what matters mathematically is just their truth, not their metaphysical necessity: mathematics does not need the concept of metaphysical necessity. Does metaphysical modality really matter outside philosophy? Even if physicists care about the physical necessity of the laws they conjecture, does it matter to physics whether physically necessary laws are also metaphysically necessary? In ordinary life, we care whether someone could have done otherwise, whether disaster could have been averted, but the kind of possibility at issue there is far more narrowly circumscribed than metaphysical possibility, by not pre-scinding from metaphysically contingent initial conditions. He could not have done otherwise because he was in chains, even though it was metaphysically contingent that he was in chains. Does “could have been” ever express metaphysical possibility when used non-philosophically?

If thought about metaphysical modality is the exclusive preserve of philosophers, so is knowledge of metaphysical modality. The epistemology of metaphysical modality tends to be treated as an isolated case. For instance, much of the discussion concerns how far, if at all, conceivability is a guide to possibility, and inconceivability to impossibility (Gendler and Hawthorne (2002) has a sample of recent contributions to this debate). The impression is that, outside philosophy, the primary cognitive role of conceiving is propaedeutic. Conceiving a hypothesis is getting it onto the table, putting it up for serious consideration as a candidate for truth. The inconceivable never even gets that far. Conceivability is certainly no good evidence for the restricted kinds of possibility we mainly care about in natural science or ordinary life. We easily conceive particles violating what are in fact physical laws, or the man without his chains. On this view, conceiving, outside philosophy, is no faculty for distinguishing truth from falsity in some domain, but rather a preliminary to any such faculty. Although there are truths and falsehoods about conceivability and inconceivability, they concern our mental capacities, whereas metaphysical modalities are supposed to be mind-independent. They are not contingent on mental capacities, because not contingent on anything (at least if we accept the principles of the modal logic S5, that the necessary is necessarily necessary and the possible necessarily possible). When philosophers present conceiving as a faculty for

distinguishing between truth and falsity in the domain of metaphysical modality, that looks suspiciously like some sort of illicit projection or unacknowledged fiction: at best, attributions of metaphysical modality would lack the cognitive status traditionally ascribed to them (compare Blackburn (1987), Craig (1985), Wright (1989), and Rosen (1990)). The apparent cognitive isolation of metaphysically modal thought makes such suspicions hard to allay. Presenting it as *sui generis* suggests that it can be surgically removed from our conceptual scheme without collateral damage. If it can, what good does it do us? In general, the postulation by philosophers of a special cognitive capacity exclusive to philosophical or quasi-philosophical thinking looks like a scam.

Humans evolved under no pressure to do philosophy. Presumably, survival and reproduction in the Stone Age depended little on philosophical prowess, dialectical skill being no more effective then than now as a seduction technique and in any case dependent on a hearer already equipped to recognize it. Any cognitive capacity we have for philosophy is a more or less accidental byproduct of other developments. Nor are psychological dispositions that are non-cognitive outside philosophy likely suddenly to become cognitive within it. We should expect the cognitive capacities used in philosophy to be cases of general cognitive capacities used in ordinary life, perhaps trained, developed, and systematically applied in various special ways, just as the cognitive capacities that we use in mathematics and natural science are rooted in more primitive cognitive capacities to perceive, imagine, correlate, reason, discuss . . . In particular, a plausible non-skeptical epistemology of metaphysical modality should subsume our capacity to discriminate metaphysical possibilities from metaphysical impossibilities under more general cognitive capacities used in ordinary life.

I will argue that the ordinary cognitive capacity to handle counterfactual conditionals carries with it the cognitive capacity to handle metaphysical modality. Section 2 illustrates with examples our cognitive use of counterfactual conditionals. Section 3 sketches an epistemology for such conditionals. Section 4 explains how they subsume metaphysical modality. Section 5 assesses the consequences for the distinction between *a priori* and *a posteriori* knowledge. Section 6 discusses some objections. Section 7 briefly raises the relation between metaphysical possibility and the restricted kinds of

possibility that seem more relevant to ordinary life. Philosophers' ascriptions of metaphysical modality are far more deeply rooted in our ordinary cognitive practices than most skeptics about it realize.

## 2

Our overall capacity for somewhat reliable thought about counterfactual possibilities is hardly surprising, for we cannot know in advance exactly which possibilities are or will be actual. We need to make contingency plans. In practice, the only way for us to be cognitively equipped to deal with the actual is by being cognitively equipped to deal with a wide variety of contingencies, most of them counterfactual. Our present task is to understand some of the more specific cognitive value to us of thinking with those conditional constructions labeled "counterfactual."

We can usefully start with a well-known example which proves the term "counterfactual conditional" misleading. As Alan Ross Anderson pointed out (1951: 37), a doctor might say:

- (1) If Jones had taken arsenic, he would have shown just exactly those symptoms which he does in fact show.

Clearly, (1) can provide abductive evidence by inference to the best explanation for its antecedent (see Edgington (2003: 23–7) for more discussion):

- (2) Jones took arsenic.

If further tests subsequently verify (2), they confirm the doctor's statement rather than in any way falsifying it or making it inappropriate. If we still call subjunctive conditionals like (1) "counterfactuals," the reason is not that they imply or presuppose the falsity of their antecedents. In what follows, we shall be just as concerned with conditional sentences such as (1) as with those whose premises are false, or believed to be so.

Of course, what (2) explains is not the trivial necessary truth that Jones shows whatever symptoms he shows. What is contingent is that Jones shows exactly those symptoms which he does in fact show – he

could have shown other symptoms, or none – and, given (1), (2) explains that contingent truth.

While (1) provides valuable empirical evidence, the corresponding indicative conditional does not (Stalnaker 1999: 71):

(1I) If Jones took arsenic, he shows just exactly those symptoms which he does in fact show.

We can safely assent to (1I) without knowing what symptoms Jones shows, since it holds whatever they are. Informally, (1) is non-trivial because it depends on a comparison between independently specified terms, the symptoms Jones would have shown if he had taken arsenic and the symptoms he does in fact show; by contrast, (1I) is trivial because it involves only a comparison of his symptoms with themselves. Thus the process of evaluating the “counterfactual” conditional requires something like two files, one for the actual situation, the other for the counterfactual situation, even if these situations turn out to coincide. No such cross-comparison of files is needed to evaluate the indicative conditional. Of course, when one evaluates an indicative conditional while disbelieving its antecedent, one must not confuse one’s file of beliefs with one’s file of judgments on the supposition of the antecedent, but that does not mean that cross-referencing from the latter file to the former can play the role it did in the counterfactual case. One logical manifestation of this difference is that any indicative conditional  $A \rightarrow @A$  is a logical truth, where  $@$  is the “actually” operator ( $@A$  is true at any given world just in case  $A$  is true at the actual world), whereas the counterfactual conditional  $A \Box \rightarrow @A$  is false if  $A$  is contingently false. For instance, I can trivially assert “If the coin landed heads, it actually landed heads,” without checking how it landed, but “If the coin had landed heads, it would have actually landed heads” is false if the coin actually landed tails, because it implies that if the coin could have landed heads, it actually did so (Williamson (2006a) has more discussion).

The sentence (1I) works differently from the non-trivial habitual:

(1H) If Jones takes arsenic, he shows just exactly those symptoms which he does in fact show.

The latter can be false when both (1) and (1I) are true, for example because Jones's symptoms are not those he would normally show on arsenic poisoning but those he would show given that he had, unusually, been fasting for the previous 72 hours, a fact the doctor took into account. Since habituals in some sense characterize "normal" cases while counterfactual conditionals can depend on abnormal features of the current case, habituals are not in general adequate substitutes for counterfactual conditionals. Of course, the truth conditions of habituals themselves involve counterfactual cases.

Since (1) constitutes empirical evidence, its truth was not guaranteed in advance. If Jones had looked suitably different, the doctor would have had to assert the opposite counterfactual conditional:

- (3) If Jones had taken arsenic, he would not have shown just exactly those symptoms which he does in fact show.

From (3) we can deduce the falsity of its antecedent. For modus ponens is generally agreed to be valid for counterfactual conditionals. Thus (2) and (3) entail:

- (4) Jones does not show just exactly those symptoms which he does in fact show.

Since (4) is obviously false, we can deny (2) given (3).

The indicative conditional corresponding to (3) is:

- (3I) If Jones took arsenic, he does not show just exactly those symptoms which he does in fact show.

To assert (3I) is like saying "If Jones took arsenic, pigs can fly." Although a very confident doctor might assert (3I), on the grounds that Jones certainly did not take arsenic, that certainty may in turn be based on confidence in (3), and therefore on the comparison of actual and counterfactual situations.

Could a Bayesian account dispense with the counterfactual conditionals in favor of conditional probabilities? Consider the simple case in which we completely trust the doctor who asserts (1). Before the doctor speaks, we are certain what symptoms Jones shows but

agnostic over the characteristic symptoms of arsenic poisoning. We want to update our probability for his having taken arsenic on evidence from the doctor, in Bayesian terms by conditionalizing on it. The doctor cannot simply tell us what probability to assign, because we may have further relevant evidence unavailable to the doctor, for example about Jones's character. We need the doctor to say something that we can use as evidence; (1) exactly fits the bill (of course, our evidence also includes the fact that the doctor asserted (1), but in the circumstances we can treat (1) itself as the relevant part of our evidence). It may even do better than a non-modal generalization such as "Jones showed exactly those symptoms which everyone who takes arsenic shows": for the symptoms may vary with bodily characteristics of the victim, and through long experience the doctor may be able to judge what symptoms Jones would have shown if he had taken arsenic without being able to articulate a suitable generalization. If he were to say "Jones showed exactly those symptoms which everyone relevantly like him who takes arsenic shows," he might easily have to do so without knowing of any instance of this contextually restricted generalization other than the one at hand; in such cases belief in the restricted generalization is epistemically based on the counterfactual conditional, not *vice versa*. Any Bayesian account depends on an adequately varied stock of propositions to act as bearers of probability, as evidence or hypotheses. Sometimes that range has to include counterfactual conditionals.

We also use the notional distinction between actual and counterfactual situations to make evaluative comparisons:

- (5) If Jones had not taken arsenic, he would have been in better shape than he now is.

Such counterfactual reflections facilitate learning from experience; one may decide never to take arsenic oneself. Formulating counterfactuals about past experience is empirically correlated with improved future performance in various tasks.<sup>1</sup>

Evidently, counterfactual conditionals give clues to causal connections. This point does not commit one to the ambitious program of

<sup>1</sup> The large empirical literature on the affective role of counterfactuals and its relation to learning from experience includes Kahneman and Tversky (1982), Roes and Olson (1993, 1995) and Byrne (2005).

analyzing causality in terms of counterfactual conditionals (Lewis 1973b), Collins, Hall, and Paul (2004)), or counterfactual conditionals in terms of causality (Jackson 1977). If the former program succeeds, all causal thinking is counterfactual thinking; if the latter succeeds, all counterfactual thinking is causal thinking. Either way, the overlap is so large that we cannot have one without much of the other. It may well be over-optimistic to expect either necessary and sufficient conditions for causal statements in counterfactual terms or necessary and sufficient conditions for counterfactual statements in causal terms. Even so, counterfactuals surely play a crucial role in our causal thinking (see Harris (2000: 118–39) and Byrne (2005: 100–28) for some empirical discussion). Only extreme skeptics deny the cognitive value of causal thought.

At a more theoretical level, claims of nomic necessity support counterfactual conditionals. If it is a law that property P implies property Q, then typically if something were to have P, it would have Q. If we can falsify the counterfactual in a specific case, perhaps by using better-established laws, we thereby falsify that claim of lawhood. We sometimes have enough evidence to establish what the result of an experiment would be without actually doing the experiment: that matters in a world of limited resources.

Counterfactual thought is deeply integrated into our empirical thought in general. Although that consideration will not deter the most dogged skeptics about our knowledge of counterfactuals, it indicates the difficulty of preventing such skepticism from generalizing implausibly far, since our beliefs about counterfactuals are so well-integrated into our general knowledge of our environment. I proceed on the assumption that we have non-trivial knowledge of counterfactuals.

### 3

In discussing the epistemology of counterfactuals, I assume no particular theory of their compositional semantics. Although I sometimes use the Stalnaker-Lewis approach for purposes of illustration and vividness, I do not assume its correctness or that of any other specific semantic account of counterfactuals, within or without the framework of possible worlds. That evasion of semantic theory might seem

dubious, since it is the semantic facts which determine what has to be known. However, we can go some way on the basis of our pretheoretical understanding of such conditionals in our native language. Moreover, the best developed formal semantic theories of counterfactuals use an apparatus of possible worlds or situations at best distantly related to our actual cognitive processing. While that does not refute such theories, which concern the truth conditions of counterfactuals, not how subjects attempt to find out whether those truth conditions obtain, it shows how indirect the relation between the semantics and the epistemology may be. When we come to fine-tune our epistemology of counterfactuals, we may need an articulated semantic theory, but at a first pass we can make do with some sketchy remarks about their epistemology while remaining as far as possible neutral over their deep semantic analysis. Although I formalize the counterfactual conditional with the usual sentence operator  $\Box \rightarrow$ , I do not assume that that exactly reflects the structure of the corresponding natural language sentences.<sup>2</sup> As for the psychological study of the processes underlying our assessment of counterfactual conditionals, it remains in a surprisingly undeveloped state, as recent authors have complained (Evans and Over 2004: 113–31).

Start with an example. You are in the mountains. As the sun melts the ice, rocks embedded in it are loosened and crash down the slope. You notice one rock slide into a bush. You wonder where it would have ended if the bush had not been there. A natural way to answer the question is by visualizing the rock sliding without the bush there, then bouncing down the slope into the lake at the bottom. Under suitable background conditions, you thereby come to know this counterfactual:

- (6) If the bush had not been there, the rock would have ended in the lake.

You could test that judgment by physically removing the bush and experimenting with similar rocks, but you know (6) even without performing such experiments. Logically, the counterfactual about the

<sup>2</sup> Lewis (1975) treats “if” in some occurrences as a restrictor on quantifiers rather than a sentential connective. This approach was generalized to all occurrences of “if” in Kratzer (1986).

past is independent of claims about future experiments (for a start, the slope is undergoing continual small changes).

Somehow, you came to know the counterfactual by using your imagination. That sounds puzzling if one conceives the imagination as unconstrained. You can imagine the rock rising vertically into the air, or looping the loop, or sticking like a limpet to the slope. What constrains imagining it one way rather than another?

You do not imagine it those other ways because your imaginative exercise is radically informed and disciplined by your perception of the rock and the slope and your sense of how nature works. The default for the imagination in its primary function may be to proceed as “realistically” as it can, subject to whatever deviations the thinker imposes by brute force: here, the absence of the bush. Thus the imagination can in principle exploit all our background knowledge in evaluating counterfactuals. Of course, how to separate background knowledge from what must be imagined away in imagining the antecedent is Goodman’s old, deep problem of cotenability (1954). For example, why don’t we bring to bear our background knowledge that the rock did not go far, and imagine another obstacle to its fall? Difficult though the problem is, it should not make us lose sight of our considerable knowledge of counterfactuals: our procedures for evaluating them cannot be too wildly misleading.

Can the imaginative exercise be regimented as a piece of reasoning? We can undoubtedly assess some counterfactuals by straightforward reasoning. For instance:

- (7) If twelve people had come to the party, more than eleven people would have come to the party.

We can deduce the consequent “More than eleven people came to the party” from the antecedent “Twelve people came to the party,” and assert (7) on that basis. Similarly, it may be suggested, we can assert (6) on the basis of inferring its consequent “The rock ended in the lake” from the premise “The bush was not there,” given auxiliary premises about the rock, the mountainside and the laws of nature.

At the level of formal logic, we have the corresponding plausible and widely accepted closure principle that, given a derivation of  $C$  from  $B_1, \dots, B_n$ , we can derive the counterfactual conditional  $A \Box \rightarrow C$  from the counterfactual conditionals  $A \Box \rightarrow B_1, \dots, A \Box \rightarrow B_n$ ; in other

words, the counterfactual consequences of a supposition  $A$  are closed under logical consequence (Lewis (1986: 132) calls this “Deduction within Conditionals”). With the uncontroversial reflexivity principle  $A \Box \rightarrow A$ , it follows that, given a derivation of  $C$  from  $A$  alone, we can derive  $A \Box \rightarrow C$  from the null set of premises.

We cannot automatically extend the closure rule to the case of auxiliary premises, for since we can derive an arbitrary conclusion  $C$  from an arbitrary premise  $A$  with  $C$  as auxiliary premise, we could then derive  $A \Box \rightarrow C$  from the auxiliary premise  $C$  alone: but that implies the invalid principle that any truth is a counterfactual consequence of any supposition whatsoever. The truth of “Napoleon lost at Waterloo” does not guarantee the truth of “If Grouchy had marched towards the sounds of gunfire, Napoleon would have lost at Waterloo.” Auxiliary premises cannot always be copied into the scope of counterfactual suppositions (this is the problem of coterminability again). Even with this caution, the treatment of the process by which we reach counterfactual judgments as inferential is problematic in several ways.

First, a technical problem: not every inference licenses us to assert the corresponding counterfactual, even when the inference is deductive and the auxiliary premises are selected appropriately. For the consequent of (1) is a logical truth (count it vacuously true if Jones shows no symptoms):

- (8) Jones shows just exactly those symptoms which he does in fact show.

Thus (8) follows from any premises, including (2), the antecedent of (1); but we cannot assert (1) on the basis of that trivial deduction alone, independently of *which* symptoms Jones does in fact show. Formally, although  $A \equiv @A$  is always a logical truth,  $B \Box \rightarrow (A \equiv @A)$  may be false. Similarly, although  $@A$  is always a logical consequence of  $A$ ,  $A \Box \rightarrow @A$  may be false. This is related to Kaplan’s (1989) point that the rule of necessitation fails in languages with terms such as “actually.” The logical truth of (8) does not guarantee the logical truth, or even truth, of (9):

- (9) It is necessary that Jones shows just exactly those symptoms which he does in fact show.

For it is contingent that Jones shows just exactly those symptoms which he does in fact show.<sup>3</sup> But let us assume that this technical problem can be solved by a restriction on the type of reasoning from antecedent to consequent that can license a counterfactual, and on the closure principle above, like the restriction on the type of reasoning that licenses the necessitation of its conclusion.

A more serious problem is that the putative reasoner may lack general-purpose cognitive access to the auxiliary premises of the putative reasoning. In particular, the folk physics needed to derive the consequents of counterfactuals such as (6) from their antecedents may be stored in the form of some analogue mechanism, perhaps embodied in a connectionist network, which the subject cannot articulate in propositional form. Normally, a subject who uses negation and derives a conclusion from some premises can at least entertain the negation of a given premise, whether or not they are willing to assert it, perhaps on the basis of the other premises and the negation of the conclusion. Our reliance on folk physics does not enable us to formulate its negation. More generally, the supposed premises may not be stored in a form that permits the normal range of inferential interactions with other beliefs, even at an unconscious level. This strains the analogy with explicit reasoning.

The third problem is epistemological. Normally, someone who believes a conclusion on the sole basis of inference from some premises knows the conclusion only if they know the premises. This principle must be applied with care, for often a thinker is aware of several inferential routes from different sets of premises to the same conclusion. For example, you believe that  $a$  and  $b$  are  $F$ ; you deduce that something is  $F$ . If you know that  $a$  is  $F$ , you may thereby come to know that something is  $F$ , even if your belief that  $b$  is  $F$  is false, and so not knowledge. Similarly, you may believe more premises than you need to draw an inductive conclusion. The principle applies only to essential premises, those that figure in all the inferences on which the relevant belief in the conclusion is based. However, folk physics is an essential standing background premise of the supposed inferences

<sup>3</sup> The phrase “does in fact show” is read throughout as inside the scope of the counterfactual conditional or modal operator, but as rigid, like “actually shows.” See Williamson (2006a) for discussion.

from antecedents to consequents of counterfactuals like (6), as usually conceived, so the epistemological maxim applies. Folk physics in this sense is a theory whose content includes the general principles by which expectations of motion, constancy, and the like are formed online in real time; it is no mere collection of memories of particular past incidents. But then presumably it is strictly speaking false: although many of its predictions are useful approximations, they are inaccurate in some circumstances; knowledge of the true laws of motion is not already wired into our brains, otherwise physics could be reduced to psychology. Since folk physics is false, it is not known. But the conclusion that no belief formed on the basis of folk physics constitutes knowledge is wildly skeptical. For folk physics is reliable enough in many circumstances to be used in the acquisition of knowledge, for example that the cricket ball will land in that field. Thus we should not conceive folk physics as a premise of that conclusion. Nor should we conceive some local fragment of folk physics as the premise. For it would be quite unmotivated to take an inferential approach overall while refusing to treat this local fragment as itself derived from the general theory of folk physics. We should conceive folk physics as a locally but not globally reliable method of belief formation, not as a premise.

If folk theories are methods of belief formation rather than specific beliefs, can they be treated as patterns of inference, for example from beliefs about the present to beliefs about the future? Represented as a universal generalization, a non-deductive pattern of inference such as abduction is represented as a falsehood, for the relevantly best explanations are not always correct. Nevertheless, we can acquire knowledge abductively because we do not rely on every abduction in relying on one; we sometimes rely on a locally truth-preserving abduction, even though abduction is not globally truth-preserving. The trouble with replacing a pattern of inference by a universal generalization is that it has us rely on all instances of the pattern simultaneously, by relying on the generalization. Even if the universal generalization is replaced by a statement of general tendencies, what we are relying on in a particular case is still inappropriately globalized. Epistemologically, folk “theories” seem to function more like patterns of inference than like general premises. That conception also solves the earlier problem about the inapplicability of logical operators to folk “theories,” since patterns of inference cannot themselves

be negated or made the antecedents of conditionals (although claims of their validity can).

Once such a liberal conception of patterns of inference is allowed, calling a process of belief formation “inferential” is no longer very informative. Just about any process with a set of beliefs (or suppositions) as input and an expanded set of beliefs (or suppositions) as output counts as “inferential.” Can we say something more informative about the imaginative exercises by which we judge counterfactuals like (6), whether or not we count them as inferential?

An attractive suggestion is that some kind of simulation is involved: the difficulty is to explain what that means. It is just a hint of an answer to say that in simulation cognitive faculties are run offline. The cognitive faculties that would be run online to evaluate  $A$  and  $B$  as free-standing sentences are run offline in the evaluation of the counterfactual conditional  $A \Box \rightarrow B$ .<sup>4</sup> This suggests that the cognition has a roughly compositional structure. Our capacity to handle  $A \Box \rightarrow B$  embeds our capacities to handle  $A$  and  $B$  separately, and our capacity to handle the counterfactual conditional operator involves a general capacity to go from capacities to handle the antecedent and the consequent separately to a capacity to handle the whole conditional. Here the capacity to handle an expression comprises more than mere linguistic understanding of it, since it involves ways of assessing its application that are not built into its meaning. But it virtually never involves a decision procedure that enables us always to determine the truth-values of every sentence in which the expression principally occurs, since we lack such decision procedures. Of course, we can sometimes take shortcuts in evaluating counterfactual conditionals. For instance, we can know that  $A \Box \rightarrow A$  is true even if we have no idea how to determine whether  $A$  is true. Nevertheless, the compositional structure just described seems more typical.

*How* do we advance from capacities to handle the antecedent and the consequent separately to a capacity to handle the whole conditional? “Offline” suggests that the most direct links with perception have been cut, but that vague negative point does not take us far.

<sup>4</sup> Matters become more complicated if  $A$  or  $B$  itself contains a counterfactual condition, as in “If she had murdered the man who would have inherited her money if she had died, she would have been sentenced to life imprisonment if she had been convicted,” but the underlying principles are the same.

Perceptual input is crucial to the evaluation of counterfactuals such as (1) and (6).

The best developed simulation theories concern our ability to simulate the mental processes of other agents (or ourselves in other circumstances), putting ourselves in their shoes, as if thinking and deciding on the basis of their beliefs and desires (see for example Davies and Stone (1995), Nichols and Stich (2003)). Such cognitive processes may well be relevant to the evaluation of counterfactuals about agents. Moreover, they would involve just the sort of constrained use of the imagination indicated above. How would Mary react if you asked to borrow her car? You could imagine her immediately shooting you, or making you her heir; you could even imagine reacting like that from her point of view, by imagining having sufficiently bizarre beliefs and desires. But you do not. Doing so would not help you determine how she really would react. Presumably, what you do is to hold fixed her actual beliefs and desires (as you take them to be just before the request); you can then imagine the request from her point of view, and think through the scenario from there. Just as with the falling rock, the imaginative exercise is richly informed and disciplined by your sense of what she is like.

How could mental simulation help us evaluate a counterfactual such as (6), which does not concern an agent? Even if you somehow put yourself in the rock's shoes, imagining first-personally being that shape, size, and hardness and bouncing down that slope, you would not be simulating the rock's reasoning and decision-making. Thinking of the rock as an agent is no help in determining its counterfactual trajectory. A more natural way to answer the question is by imagining third-personally the rock falling as it would visually appear from your actual present spatial position; you thereby avoid the complex process of adjusting your current visual perspective to the viewpoint of the rock. Is that to simulate the mental states of an observer watching the rock fall from your present position?<sup>5</sup> By itself, that suggestion explains little. For how do we know what to simulate the observer seeing next?

That question is not unanswerable. For we have various propensities to form expectations about what happens next: for example, to

<sup>5</sup> See Goldman (1992: 24), discussed by Nichols, Stich, Leslie, and Klein (1996: 53–9).

project the trajectories of nearby moving bodies into the immediate future (otherwise we could not catch balls). Perhaps we simulate the initial movement of the rock in the absence of the bush, form an expectation as to where it goes next, feed the expected movement back into the simulation as seen by the observer, form a further expectation as to its subsequent movement, feed that back into the simulation, and so on. If our expectations in such matters are approximately correct in a range of ordinary cases, such a process is cognitively worthwhile. The very natural laws and causal tendencies our expectations roughly track also help to determine which counterfactual conditionals really hold. Thus some reliability in the assessment of counterfactuals is achieved.

However, talk of simulating the mental states of an observer may suggest that the presence of the observer is part of the content of the simulation. That does not fit our evaluation of counterfactuals. Consider:

(10) If there had been a tree on this spot a million years ago, nobody would have known.

Even if we visually imagine a tree on this spot a million years ago, we do not automatically reject (10) because we envisage an observer of the tree. We may imagine the tree as having a certain visual appearance from a certain viewpoint, but that is not to say that we imagine it as appearing to someone at that viewpoint. For example, if we imagine the sun as shining from behind that viewpoint, by imagining the tree's shadow stretching back from the tree, we are not obliged to imagine either the observer's shadow stretching towards the tree or the observer as perfectly transparent.<sup>6</sup> Nor, when we consider (10),

<sup>6</sup> The question is of course related to Berkeley's claim that we cannot imagine an unseen object. For discussion see Williams (1966), Peacocke (1985) and Currie (1995b: 36–7). Gaut (2006: 116–21) describes the role of art in facilitating the evaluation of counterfactuals by means of the imagination. He disavows commitment to the view, which he credits to Currie (1995a) (ch. 5), that "imagination is a kind of 'offline' running of cognitive processes, and that this is a source of knowledge of psychological states," appealing instead to the tradition of Vico and Weber, on which the relevant role of imagination is in *verstehen*, in understanding oneself and others (Gaut 2006: 121). However, it is doubtful that this tradition can (or wants to) explain knowledge of counterfactuals that do not concern mental states.

are we asking whether if we had believed that there was a tree on this spot a million years ago, we would have believed that nobody knew.<sup>7</sup> It is better not to regard the content of the simulation as referring to anything specifically *mental* at all. It is just that visual imagining reuses offline some of the very same cognitive resources that visual perceiving uses online.

Of course, for many counterfactuals the relevant expectations are not hardwired into us in the way that those concerning the trajectories of fast-moving objects around us may need to be. Our knowledge that if a British general election had been called in 1948 the Communists would not have won may depend on an offline use of our capacity to predict political events. Still, where our more sophisticated capacities to predict the future are reliable, so should be corresponding counterfactual judgments. In these cases too, simulating the mental states of an imaginary observer seems unnecessary.

The offline use of expectation-forming capacities to judge counterfactuals corresponds to the widespread picture of the semantic evaluation of those conditionals as “rolling back” history to shortly before the time of the antecedent, modifying its course by stipulating the truth of the antecedent and then rolling history forward again according to patterns of development as close as possible to the normal ones to test the truth of the consequent (compare Lewis (1979)).

The use of expectation-forming capacities may in effect impose a partial solution to Goodman’s problem of cotenability, since they do not operate on information about what happened after the time treated as present. In this respect indicative conditionals are evaluated differently: if I had climbed a mountain yesterday I would remember

<sup>7</sup> A similar problem arises for what is sometimes called the Ramsey Test for conditionals, on which one simulates belief in the antecedent and asks whether one then believes the consequent. Goldman (1992: 24) writes “When considering the truth value of ‘If X were the case, then Y would obtain,’ a reasoner feigns a belief in X and reasons about Y under that pretence.” What Ramsey himself says is that when people “are fixing their degrees of belief in  $q$  given  $p$ ” they “are adding  $p$  hypothetically to their stock of knowledge and arguing on that basis about  $q$ ” (1978: 143), but he specifically warns that “the degree of belief in  $q$  given  $p$ ” does not mean the degree of belief “which the subject would have in  $q$  if he knew  $p$ , or that which he ought to have” (1978: 82; variables interchanged). Of course, conditional probabilities bear more directly on indicative than on subjunctive conditionals.

it today, but if I did climb a mountain yesterday I do not remember it today. The known fact that I do not remember climbing a mountain yesterday is retained under the indicative but not the counterfactual supposition.

Our offline use of expectation-forming capacities to unroll a counterfactual history from the imagined initial conditions does not explain why we imagine the initial conditions in one way rather than another – for instance, why we do not imagine a wall in place of the bush. Very often, no alternative occurs to us, but that does not mean that the way we go adds nothing to the given antecedent. We seem to have a prereflective tendency to minimum alteration in imagining counterfactual alternatives to actuality, reminiscent of the role that similarity between possible worlds plays in the Lewis-Stalnaker semantics.

Of course, not all counterfactual conditionals can be evaluated by the rolling back method, since the antecedent need not concern a particular time: in evaluating the claim that space-time has ten dimensions, a scientist can sensibly ask whether if it were true the actually observed phenomena would have occurred. Explicit reasoning may play a much larger role in the evaluation of such conditionals.

Reasoning and prediction do not exhaust our capacity to evaluate counterfactuals. If twelve people had come to the party, would it have been a large party? To answer, one does not imagine a party of twelve people and then predict what would happen next. The question is whether twelve people would have constituted a large party, not whether they would have caused one. Nor is the process of answering best conceived as purely inferential, if one has no special antecedent beliefs as to how many people constitute a large party, any more than the judgment whether the party is large is purely inferential when made at the party. Rather, in both cases one must make a new judgment, even though it is informed by what one already believes or imagines about the party. To call the new judgment “inferential” simply because it is not made independently of all the thinker’s prior beliefs or suppositions is to stretch the term “inferential” beyond its useful span. At any rate, the judgment cannot be derived from the prior beliefs or suppositions purely by the application of general rules of inference. For example, even if you have the prior belief that a party is large if and only if it is larger than the average size of a party, in order to apply it to the case at hand you also need to have a belief

as to what the average size of a party is; if you have no prior belief as to that, and must form one by inference, an implausible regress threatens, for you do not have the statistics of parties in your head. Similarly, if you try to judge whether this party is large by projecting inductively from previous judgments as to whether parties were large, that only pushes the question back to how those previous judgments were made.

In general, our capacity to evaluate counterfactuals recruits *all* our cognitive capacities to evaluate sentences. A quick argument for this uses the assumption that a counterfactual with a true antecedent has the same truth-value as its consequent, for then any sentence  $A$  is logically equivalent to  $T \rightarrow A$ , where  $T$  is a trivial tautology; so any non-logical cognitive work needed to evaluate  $A$  is also needed to evaluate the counterfactual  $T \rightarrow A$ .<sup>8</sup> For if we could evaluate that counterfactual without doing the non-logical work, we could also evaluate  $A$  without doing it, by first evaluating the counterfactual, then deriving its equivalence to  $A$  and finally extending the evaluation of the former to the latter. Any logical work needed to evaluate  $A$  will also be needed to evaluate  $T \rightarrow A$  when  $T$  is chosen to be irrelevant to  $A$ .

There is no uniform epistemology of counterfactual conditionals. In particular, imaginative simulation is neither always necessary nor always sufficient for their evaluation, even when they can be evaluated. Nevertheless, it is the most distinctive cognitive feature of the process of evaluating them, because it is so much more useful for counterfactuals than for most non-counterfactual contents, whereas reasoning, perception, and testimony are not generally more useful for counterfactuals than for non-counterfactual contents.

We can still schematize a typical overall process of evaluating a counterfactual conditional thus: one supposes the antecedent and

<sup>8</sup> Lewis (1986: 26–31) defends the assumption; Nozick (1981: 176) rejects it to make the fourth condition in his analysis of knowledge non-trivial. Bennett (2003: 239–40) also rejects it. The point can be made independently of that assumption, using the rigidifying “actually” operator  $@$ . For  $@A$  entails  $B \rightarrow @A$  for any  $B$  and therefore  $T \rightarrow @A$  in particular; conversely,  $T \rightarrow @A$  entails  $@A$  by modus ponens. Since  $A$  is logically equivalent to  $@A$ , it is logically equivalent to  $T \rightarrow @A$ . Thus any cognitive capacities needed to assess  $A$  will also be needed to assess the more complex  $T \rightarrow @A$  (modulo those needed to recognize the equivalence).

develops the supposition, adding further judgments within the supposition by reasoning, offline predictive mechanisms, and other offline judgments. The imagining may but need not be perceptual imagining. All of one's background knowledge and beliefs are available from within the scope of the supposition as a description of one's actual circumstances for the purposes of comparison with the counterfactual circumstances (if we know  $B$ , we can infer  $A \Box \rightarrow @B$  for any  $A$ ; in this respect the development differs from that of the antecedent of an indicative conditional). Some but not all of one's background knowledge and beliefs are also available within the scope of the supposition as a description of the counterfactual circumstances, according to complex criteria (the problem of cotenability). To a first approximation: one asserts the counterfactual conditional if and only if the development eventually leads one to add the consequent.

An over-simplification in that account is that one develops the initial supposition only once. In fact, if one finds various different ways of imagining the antecedent equally good, one may try developing several of them, to test whether they all yield the consequent. For example, if in considering (10) one initially imagines a palm tree, one does not immediately judge that if there had been a tree on this spot a million years ago it would have been a palm tree, because one knows that one can equally easily imagine a fir tree. One repeats the thought experiment. Robustness in the result under such minor perturbations supports a higher degree of confidence.

What happens if the counterfactual development of the antecedent  $A$  does not robustly yield the consequent  $C$ ? We do not always deny  $A \Box \rightarrow C$ , for several reasons. First, if  $C$  has not emerged after a given period of development the question remains whether it will emerge in the course of further development, for lines of reasoning can be continued indefinitely from any given premise. To reach a negative conclusion, one must in effect judge that if the consequent were ever going to emerge it would have done so by now. For example, one may have been smoothly fleshing out a scenario incompatible with the consequent with no hint of difficulty. Second, even if one is confident that  $C$  will not robustly emerge from the development, one may suspect that the reason is one's ignorance of relevant background conditions rather than the lack of a counterfactual connection between  $A$  and  $C$  ("If I were to follow that path, it would lead me out of the forest"). Thus one may remain agnostic over  $A \Box \rightarrow C$ .

The case for denying  $A \rightarrow C$  is usually strongest when the counterfactual development of  $A$  yields  $\neg C$ . Then one asserts the opposite counterfactual,  $A \rightarrow \neg C$ . The default is to deny a counterfactual if one asserts the opposite counterfactual, moving from  $A \rightarrow \neg C$  to  $\neg(A \rightarrow C)$ . The move is defeasible; sometimes one must accept opposite counterfactuals together. For example, deductive closure generates both  $(B \ \& \ \neg B) \rightarrow B$  and  $(B \ \& \ \neg B) \rightarrow \neg B$ . Normally, if the counterfactual development of  $A$  robustly yields  $\neg C$  and robustly fails to yield  $C$  then one denies  $A \rightarrow C$ , but even this connection is defeasible, since one may still suspect that  $C$  (as well as  $\neg C$ ) would emerge given more complex reasoning or further background information.

Sometimes a counterfactual antecedent is manifestly neutral between contradictory consequents: consider “If the coin had been tossed it would have come up heads” and “If the coin had been tossed it would have come up tails.” In such cases one will clearly never be in a position to assert one conditional, and thus will never be in a position to use it as a basis for denying the opposite conditional. Whether the symmetry permits one to deny both conditionals is controversial.<sup>9</sup>

The epistemological asymmetry between asserting and denying a counterfactual conditional resembles an epistemological asymmetry in practice between asserting and denying many existential claims. If I find snakes in Iceland, without too much fuss I can assert that there are snakes in Iceland. If I fail to find snakes in Iceland, I cannot deny that there are snakes in Iceland without some implicit or explicit assessment of the thoroughness of my search: if there were snakes in Iceland, would I have found some by now? But we are capable of making such assessments, and sometimes are in a position to deny such existential claims. Similarly, if I find a counterfactual connection

<sup>9</sup> On the Lewis semantics, both  $A \rightarrow C$  and  $A \rightarrow \neg C$  are false when there is a tie for the closest  $A$  worlds to the actual world and some but not all of the joint winners are  $C$  worlds. Thus we may truly assert both  $\neg(A \rightarrow C)$  and  $\neg(A \rightarrow \neg C)$ . On the Stalnaker semantics, “Conditional Excluded Middle”  $(A \rightarrow C) \vee (A \rightarrow \neg C)$  is a logical law, because a unique  $A$  world must be selected, but sometimes neither disjunct is determinately true, because it is indeterminate which  $A$  world is selected. In such cases, neither disjunct is determinately false, so we cannot truly assert either  $\neg(A \rightarrow C)$  or  $\neg(A \rightarrow \neg C)$ ; we must simply reject both  $A \rightarrow C$  and  $A \rightarrow \neg C$  as not definitely true.

between **A** and **C** (my counterfactual development of **A** robustly yields **C**) without too much fuss I can assert  $A \square \rightarrow C$ . If I fail to find a counterfactual connection between **A** and **C** (my counterfactual development of **A** does not robustly yield **C**), I cannot deny  $A \square \rightarrow C$  without some implicit or explicit assessment of the thoroughness of my search: if there were a counterfactual connection, would I have found it by now? But we are capable of making such assessments, and sometimes are in a position to deny counterfactual conditionals.

For both assertions and denials of counterfactuals, the reliability of our cognitive faculties in their online applications across a wide range of possible circumstances induces reliability in their offline applications too. Offline reliability is achieved even with respect to counterfactual circumstances in which we would not be around to apply those faculties (“If there had been no sentient beings . . .”), for online reliability is often best achieved by tracking robust underlying trends (in nature, in logic, . . .) that hold irrespective of the presence of an observer.

The preceding remarks are the merest sketch of an epistemology of counterfactuals. Nevertheless, they will serve for purposes of orientation in what follows.

Despite its discipline, our imaginative evaluation of counterfactual conditionals is manifestly fallible. We can easily misjudge their truth-values, through background ignorance or error, and distortions of judgment. But such fallibility is the common lot of human cognition. Our use of the imagination in evaluating counterfactuals is moderately reliable and practically indispensable. Rather than cave in to skepticism, we should admit that our methods sometimes yield knowledge of counterfactuals.

## 4

How does the epistemology of counterfactual conditionals bear on the epistemology of metaphysical modality? We can approach this question by formulating two plausible constraints on the relation between counterfactual conditionals and metaphysical modalities. Henceforth, “necessary” and “possible” will be used for the metaphysical modalities unless otherwise stated.

First, the strict conditional implies the counterfactual conditional:

$$\text{NECESSITY } \square(A \rightarrow B) \rightarrow (A \square \rightarrow B)$$

Suppose that  $A$  could not have held without  $B$  holding too; then if  $A$  had held,  $B$  would also have held. In terms of possible worlds semantics for these operators along the lines of Lewis (1973) or Stalnaker (1968): if all  $A$  worlds are  $B$  worlds, then any closest  $A$  worlds are  $B$  worlds. More precisely, if all  $A$  worlds are  $B$  worlds, then either there are no  $A$  worlds or there is an  $A$  world such that any  $A$  world at least as close as it is to the actual world is a  $B$  world.

Second, the counterfactual conditional transmits possibility:

$$\text{POSSIBILITY } (A \square \rightarrow B) \rightarrow (\Diamond A \rightarrow \Diamond B)$$

Suppose that if  $A$  had held,  $B$  would also have held; then if  $A$  could have held,  $B$  could also have held. In terms of worlds: if any closest  $A$  worlds are  $B$  worlds, and there are  $A$  worlds, then there are also  $B$  worlds. More precisely, if either there are no  $A$  worlds or there is an  $A$  world such that any  $A$  world at least as close as it is to the actual world is a  $B$  world, then if there is an  $A$  world there is also a  $B$  world.

Together, NECESSITY and POSSIBILITY sandwich the counterfactual conditional between two modal conditions. But they do not squeeze it very tight, for  $\Diamond A \rightarrow \Diamond B$  is much weaker than  $\square(A \rightarrow B)$ : although the latter entails the former in any normal modal logic, the former is true and the latter false whenever  $B$  is possible without being a necessary consequence of  $A$ , for example when  $A$  and  $B$  are modally independent.

Although NECESSITY and POSSIBILITY determine no necessary and sufficient condition for the counterfactual conditional in terms of necessity and possibility, they yield necessary and sufficient conditions for necessity and possibility in terms of the counterfactual conditional.

We argue thus. Let  $\perp$  be a contradiction. As a special case of NECESSITY:

$$(11) \quad \square(\neg A \rightarrow \perp) \rightarrow (\neg A \square \rightarrow \perp)$$

By elementary modal logic (specifically, the weakest normal modal logic K, used throughout), since a truth-functional consequence of something necessary is itself necessary:

$$(12) \quad \Box A \rightarrow \Box(\neg A \rightarrow \perp)$$

From (11) and (12) by transitivity of the material conditional:

$$(13) \quad \Box A \rightarrow (\neg A \Box \rightarrow \perp)$$

Similarly, as a special case of POSSIBILITY:

$$(14) \quad (\neg A \Box \rightarrow \perp) \rightarrow (\Diamond \neg A \rightarrow \Diamond \perp)$$

By elementary modal logic, since the possibility of a contradiction is itself inconsistent, and necessity is the dual of possibility (being necessary is equivalent to having an impossible negation):

$$(15) \quad (\Diamond \neg A \rightarrow \Diamond \perp) \rightarrow \Box A$$

From (14) and (15) by transitivity:

$$(16) \quad (\neg A \Box \rightarrow \perp) \rightarrow \Box A$$

Putting (13) and (16) together:

$$(17) \quad \Box A \equiv (\neg A \Box \rightarrow \perp)$$

The necessary is that whose negation counterfactually implies a contradiction. Since possibility is the dual of necessity (being possible is equivalent to having an unnecessary negation), (17) yields a corresponding necessary and sufficient condition for possibility, once a double negation in the antecedent of the counterfactual has been eliminated.

$$(18) \quad \Diamond A \equiv \neg(A \Box \rightarrow \perp)$$

The impossible is that which counterfactually implies a contradiction; the possible is that which does not. In (17) and (18), the difference

between necessity and possibility lies simply in the scope of negation.

Without assuming a specific framework for the semantics of counterfactuals (in particular, that of possible worlds), we can give a simple semantic rationale for (17) and (18), based on the idea of vacuous truth. That some true counterfactuals have impossible antecedents is clear, for otherwise  $A \rightarrow A$  would fail when  $A$  was impossible. Make two widely accepted assumptions about the distinction between vacuous and non-vacuous truth: (a)  $B \rightarrow C$  is vacuously true if and only if  $B$  is impossible (this is almost a definition of “vacuously” for counterfactuals); (b)  $B \rightarrow C$  is non-vacuously true only if  $C$  is possible. The truth of (17) and (18) follows, given normal modal reasoning. If  $\Box A$  is true, then  $\neg A$  is impossible, so by (a)  $\neg A \rightarrow \perp$  is vacuously true; conversely, if  $\neg A \rightarrow \perp$  is true, then by (b) it is vacuously true, so by (a)  $\neg A$  is impossible, so  $\Box A$  is true. Similarly, if  $\Diamond A$  is true, then  $A$  is not impossible, so by (a)  $A \rightarrow \perp$  is not vacuously true, and by (b) not non-vacuously true, so  $\neg(A \rightarrow \perp)$  is true; if  $\Diamond A$  is not true, then  $A$  is impossible, so by (a)  $A \rightarrow \perp$  is vacuously true, so  $\neg(A \rightarrow \perp)$  is not true.

Given that the equivalences (17) and (18) and their necessities are logically true, metaphysically modal thinking is logically equivalent to a special case of counterfactual thinking. Thus, modulo the implicit recognition of this equivalence, the epistemology of metaphysically modal thinking is tantamount to a special case of the epistemology of counterfactual thinking. Whoever has what it takes to understand the counterfactual conditional and the elementary logical auxiliaries  $\neg$  and  $\perp$  has what it takes to understand possibility and necessity operators.

The definability of necessity and possibility in terms of counterfactual conditionals was recognized long ago. It is easy to show from the closure and reflexivity principles for  $\rightarrow$  in Section 3 that  $A \rightarrow \perp$  is logically equivalent to  $A \rightarrow \neg A$ . Thus (17) and (18) generate two new equivalences:

$$(19) \quad \Box A \equiv (\neg A \rightarrow A)$$
$$(20) \quad \Diamond A \equiv \neg(A \rightarrow \neg A)$$

The necessary is that which is counterfactually implied by its own negation; the possible is that which does not counterfactually imply

its own negation. Stalnaker (1968) used (19) and (20) to define necessity and possibility, although his reading of the conditional (with a different notation) was not exclusively counterfactual. Lewis (1973a: 25) used (17) and (18) themselves to define necessity and possibility in terms of the counterfactual conditional. However, such definitions seem to have been treated as convenient notational economies, their potential philosophical significance unnoticed (Hill (2006) is a recent exception).

If we permit ourselves to quantify into sentence position (“propositional quantification”), we can formulate another pair of variants on (17) and (18) that may improve our feel for what is going on.<sup>10</sup> On elementary assumptions about the logic of such quantifiers and of the counterfactual conditional,  $\neg A \rightarrow A$  is provably equivalent to  $\forall p (p \rightarrow A)$ : something is counterfactually implied by its negation if and only if it is counterfactually implied by everything. Thus (19) and (20) generate these equivalences too:

(21)  $\Box A \equiv \forall p (p \rightarrow A)$   
 (22)  $\Diamond A \equiv \exists p \neg(p \rightarrow \neg A)$

According to (21), something is necessary if and only if whatever were the case, it would still be the case (see also Lewis 1986: 23). That is a natural way of explaining informally what metaphysically necessity is. According to (22), something is possible if and only if it is not such that it would fail in every eventuality.

We can plausibly treat NECESSITY and POSSIBILITY as axiom schemas of a joint logic of modality and counterfactuals, susceptible in the usual way to necessitation and the analogous closure principles for counterfactuals. Then (17)–(22) will be theorems, and susceptible to the same rules. Consequently, the result of substituting the left-hand for the right-hand side of any of these biconditionals or *vice versa* anywhere in any formula built up out of atomic sentences using

<sup>10</sup> This quantification into sentence position need not be understood substitutionally. In purely modal contexts it can be modeled as quantification over all sets of possible worlds, even if not all of them are intensions of sentences that form the supposed substitution class, although this modeling presumably fails for hyperintensional contexts such as epistemic ones. A more faithful semantics for it might use non-substitutional quantification into sentence position in the metalanguage. Such subtleties are inessential for present purposes.

the modal operators, the counterfactual conditional, and truth-functors will be logically equivalent to the original (see also Appendix 1; the restrictions on necessitation and the closure principles discussed there are not relevant here).

Since the right-hand sides of (17), (19), and (21) are not strictly synonymous with each other, given the differences in their semantic structure, they are not all strictly synonymous with  $\Box A$ . Similarly, since the right-hand sides of (18), (20), and (22) are not strictly synonymous with each other, they are not all strictly synonymous with  $\Diamond A$ . Indeed, we have no sufficient reason to regard any of the equivalences as strict synonymies. That detracts little from their philosophical significance, for failure of strict synonymy does not imply failure of logical equivalence. The main philosophical concerns about possibility and necessity apply equally to anything logically equivalent to possibility or necessity. A non-modal analogy:  $\neg A$  is logically equivalent to  $A \rightarrow \perp$ , but presumably they are not strictly synonymous; nevertheless, once we have established that a creature can handle  $\rightarrow$  and  $\perp$ , we have established that it can handle something logically equivalent to negation, which answers the most interesting questions about its ability to handle negation. We should find the mutual equivalence of (17), (19), and (21), and of (18), (20), and (22) reassuring, for it shows the robustness of the modal notions definable from the counterfactual conditional, somewhat as the equivalence of the various proposed definitions of “computable function” showed the robustness of that notion.

If we treat (17) and (18) like definitions of  $\Box$  and  $\Diamond$  for logical purposes, and assume some elementary principles of the logic of counterfactuals, then we can establish the main principles of elementary modal logic for  $\Box$  and  $\Diamond$ . For example, we can show that what follows from necessary premises is itself necessary. Given that counterfactual conditionals obey modus ponens (or even weaker assumptions), we can show that what is necessary is the case. We can also check that the principles NECESSITY and POSSIBILITY, which we used to establish (17) and (18), do indeed hold under the latter characterizations of necessity and possibility. Under much stronger assumptions about the logic of the counterfactual conditional, we can also establish much stronger principles of modal logic, such as the S5 principle that what is possible is necessarily possible. Such connections extend to quantified modal logic. The logic of counterfactual

conditionals smoothly generates the logic of the modal operators (Appendix 1 gives technical details).

In particular, the proposed conception of modality makes quantification into the scope of modal operators tantamount to a special case of quantification into counterfactual contexts, as in (23) and (24):

- (23) Everyone who would have benefited if the measure had passed voted for it.
- (24) Where would the rock have landed if the bush had not been there?

Thus challenges to the intelligibility of claims of *de re* necessity are tantamount to challenges to the intelligibility of counterfactuals such as (23) and (24). But (23) and (24) are evidently intelligible.

Other properties of metaphysical modality follow from corresponding properties of counterfactual conditionals. For instance, if this is identical with that then what would have been the case of this in given counterfactual circumstances is what would have been the case of that in those circumstances; thus  $x = y$  and the triviality  $x \neq y \rightarrow x \neq y$  yield  $x \neq y \rightarrow x \neq x$ ; hence  $x = y$  entails  $x \neq y \rightarrow x \neq x$ ; since  $x \neq x$  entails  $\perp$ ,  $x = y$  entails  $x \neq y \rightarrow \perp$  and therefore  $\neg\Diamond x \neq y$ , which is a form of the law of the necessity of identity.<sup>11</sup> Again, consider the Kripkean conception of the essentiality of origin, on which, very roughly, an object could not have originated otherwise than it actually did. It follows from the plausible assumption that if something in any circumstance had originated otherwise than a given object actually did, it would not have been that very object. By contrast, objects could easily have ended otherwise than they actually did. That temporal asymmetry seems to be related to more general temporal asymmetries in the evaluation of counterfactual conditionals by the “rolling back” procedure mentioned above, which involves holding fixed an initial segment of the past but not a final segment of the future.

<sup>11</sup> In his 1961 dissertation, Dagfin Føllesdal was already clear that problems of quantifying in and substitution of coreferential terms arise for counterfactual conditionals just as they do for modal operators, although the direct connection he envisaged was through an analysis of counterfactuals in terms of natural necessity (2004: 14, 99).

Given (17) and (18), we should expect the epistemology of metaphysical modality to be a special case of the epistemology of counterfactuals. Despite the non-synonymy of the two sides, our cognitive capacity to evaluate the counterfactual conditionals gives us exactly what we need to evaluate the corresponding modal claims too. The idea that nevertheless we evaluate them by some quite different means is highly fanciful, since it indicates a bizarre lack of cognitive economy and has no plausible explanation of where the alternative cognitive resources might come from. Furthermore, as we shall see, characteristic features of the epistemology of modality are well explained by subsumption under corresponding features of the epistemology of counterfactuals. Far from being *sui generis*, the capacity to handle metaphysical modality is an “accidental” byproduct of the cognitive mechanisms that provide our capacity to handle counterfactual conditionals. Since our capacity for modal thinking cannot be isolated from our capacity for ordinary thinking about the natural world, which involves counterfactual thinking, skeptics about metaphysical modality cannot excise it from our conceptual scheme without loss to ordinary thought about the natural world, for the former is implicit in the latter.

A useful comparison is with the relation between logical consequence and logical truth. Consider some agents who reason in simple ways about themselves and their environment, perhaps using rules of inference formalizable in a Gentzen-style natural deduction calculus, perhaps in some less sophisticated way. The practical value of their reasoning skill is that they can move from ordinary empirical premises to ordinary empirical conclusions in ways that always preserve truth, thereby extending their knowledge of mundane matters (see Schechter 2006 for discussion). In doing so, they need never use logically true sentences. Nevertheless, the cognitive capacity that enables them to make these transitions between empirical sentences also enables them, as a special case, an “accidental” byproduct, to deduce logical truths from the null set of premises. Highly artificial moves would be needed to block these bonus deductions; such *ad hoc* restrictions would come at the price of extra computational complexity for no practical gain. Likewise at the semantic level: the simplest compositional semantics that enables us to negate and conjoin empirical sentences also enables us to formulate logical truths and false-

hoods, even if we have hitherto lacked any interest in doing so. By good fortune, everything is already in place for the logician to evaluate logical truths and falsehoods (at least in first-order logic, since it is complete). The philosopher's position with respect to metaphysical modality is not utterly different.

Discussions of the epistemology of modality often focus on imaginability or conceivability as a test of possibility while ignoring the role of the imagination in the assessment of mundane counterfactuals. In doing so, they omit the appropriate context for understanding the relation between modality and the imagination. For instance, scorn is easily poured on imagination as a test of possibility: it is imaginable but not possible that water does not contain oxygen, except in artificial senses of “imaginable” that come apart from possibility in other ways, and so on. Imagination can be made to look cognitively worthless. Once we recall its fallible but vital role in evaluating counterfactual conditionals, we should be more open to the idea that it plays such a role in evaluating claims of possibility and necessity. At the very least, we cannot expect an adequate account of the role of imagination in the epistemology of modality if we lack an adequate account of its role in the epistemology of counterfactuals.

On the rough sketch in Section 3, we assert  $A \Box \rightarrow B$  when our counterfactual development of the supposition  $A$  robustly yields  $B$ ; we deny  $A \Box \rightarrow B$  when our counterfactual development of  $A$  does not robustly yield  $B$  (and we do not attribute the failure to a defect in our search). Correspondingly, by (17), we assert  $\Box A$  when our counterfactual development of the supposition  $\neg A$  robustly yields a contradiction; we deny  $\Box A$  when our counterfactual development of  $\neg A$  does not robustly yield a contradiction (and we do not attribute the failure to a defect in our search). Similarly, by (18), we assert  $\Diamond A$  when our counterfactual development of the supposition  $A$  does not robustly yield a contradiction (and we do not attribute the failure to a defect in our search); we deny  $\Diamond A$  when our counterfactual development of  $A$  robustly yields a contradiction. Thus our fallible imaginative evaluation of counterfactuals has a conceivability test for possibility and an inconceivability test for impossibility built in as fallible special cases.

Such conceivability and inconceivability will be subject to the same constraints, whatever they are, as counterfactual conditionals

in general, concerning which parts of our background information are held fixed. If we know enough chemistry, our counterfactual development of the supposition that gold is the element with atomic number 79 will generate a contradiction. The reason is not simply that we know that gold is the element with atomic number 79, for we can and must vary some items of our knowledge under counterfactual suppositions. Rather, part of the general way we develop counterfactual suppositions is to hold such constitutive facts fixed.

A nuanced account of our handling of counterfactuals is likely to predict that we are more reliable in evaluating some kinds than others. For example, we may well be more reliable in evaluating counterfactuals whose antecedents involve small departures from the actual world than in evaluating those whose antecedents involve much larger departures. We may be correspondingly more reliable in evaluating the possibility of everyday scenarios than of “far-out” ones, and extra caution may be called for in the latter case. At the limit, actuality is often the best argument for possibility. But current philosophical practice already shows some sensitivity to such considerations. Many philosophers are more confident in their judgments about more or less realistic thought experiments in epistemology and moral philosophy than about more radically strange ones in metaphysics. More explicit consideration of the link between modal thought and counterfactual thought may lead to further refinements of our practice.

The considerations of this chapter will not resolve every fraught dispute about metaphysical modality, such as whether zombies (unconscious physical duplicates of us) are possible. For suppose that the source of such a dispute really is the failure of our usual methods for resolving modal issues to issue a clear verdict in the case at hand – rather than, say, the unsolvability of a non-modal problem about the nature of consciousness. Then since the present account characterizes our usual method, rather than proposing an alternative, it cannot be expected to resolve the dispute. For all that has been argued here, we may in many cases be incapable of coming to know whether a given hypothesis is metaphysically possible. Philosophical controversy will naturally make the unclear cases salient. That should not blind us to the wide range of clear cases (talking donkeys are possible). General skepticism in the epistemology of metaphysical

modality without general skepticism in the epistemology of counterfactuals is unmotivated. The use of imagination to evaluate philosophical claims of possibility and necessity is just as legitimate in principle, and sometimes just as effective in practice, as is its use to evaluate mundane counterfactuals.

## 5

What does the envisaged assimilation of modality to counterfactual conditionals imply for the status of modal judgments as knowable *a priori* or only *a posteriori*?

Some counterfactual conditionals look like paradigms of *a priori* knowability: for example (7), whose consequent is a straightforward deductive consequence of its antecedent. Others look like paradigms of what can be known only *a posteriori*: for example, that if I had searched in my pocket five minutes ago I would have found a coin. But those are easy cases.

Standard discussions of the *a priori* distinguish between two roles that experience plays in cognition, one *evidential*, one *enabling*. Experience is held to play an evidential role in my visual knowledge that this shirt is green, but a merely enabling role in my knowledge that all green things are colored: I needed it only to acquire the concepts *green* and *colored*, without which I could not even raise the question whether all green things are colored. Knowing *a priori* is supposed to be incompatible with an evidential role for experience, or at least with an evidential role for sense experience, so my knowledge that this shirt is green is not *a priori*. By contrast, knowing *a priori* is supposed to be compatible with an enabling role for experience, so my knowledge that all green things are colored can still be *a priori*. However, in our imagination-based knowledge of counterfactuals, sense experience can play a role that is neither strictly evidential nor purely enabling. For, even without surviving as part of our total evidence, it can mold our habits of imagination and judgment in ways that go far beyond a merely enabling role.

Here is an example. I acquire the words “inch” and “centimeter” independently of each other. Through sense experience, I learn to make naked eye judgments of distances in inches or centimeters with moderate reliability. When things go well, such judgments amount

to knowledge: *a posteriori* knowledge, of course. For example, I know *a posteriori* that two marks in front of me are at most two inches apart. Now I deploy the same faculty offline to make a counterfactual judgment:

(25) If two marks had been nine inches apart, they would have been at least nineteen centimeters apart.

In judging (25), I do not use a conversion ratio between inches and centimeters to make a calculation. In the example I know no such ratio. Rather, I visually imagine two marks nine inches apart, and use my ability to judge distances in centimeters visually offline to judge under the counterfactual supposition that they are at least nineteen centimeters apart. With this large margin for error, my judgment is reliable. Thus I know (25). Do I know it *a priori* or *a posteriori*? Sense experience plays no direct evidential role in my judgment. I do not consciously or unconsciously recall memories of distances encountered in perception, nor do I deduce (25) from general premises I have inductively or abductively gathered from experience: Section 3 noted obstacles to assimilating such patterns of counterfactual judgment to the use of general premises. Nevertheless, the causal role of past sense experience in my judgment of (25) far exceeds enabling me to grasp the concepts relevant to (25); the weakness of the conditions for concept possession was noted in the previous chapter. Someone could easily have enough sense experience to understand (25) without being reliable enough in their judgments of distance to know (25). Nor is the role of past experience in the judgment of (25) purely enabling in some other way, for example by acquainting me with a logical argument for (25). It is more directly implicated than that. Whether my belief in (25) constitutes knowledge is highly sensitive to the accuracy or otherwise of the empirical information about lengths (in each unit) on which I relied when calibrating my judgments of length (in each unit). I know (25) only if my offline application of the concepts of an inch and a centimeter was sufficiently skilful. Whether I am justified in believing (25) likewise depends on how skilful I am in making such judgments. My possession of the appropriate skills depends constitutively, not just causally, on past experience for the calibration of my judgments of length in those units. If the calibration is correct by a lucky accident,

despite massive errors in the relevant past beliefs about length, I lack the required skill.<sup>12</sup>

If we knew counterfactual conditionals by purely *a priori* inference from the antecedent and background premises to the conclusion, our knowledge might count as *a priori* if we knew all the background premises *a priori*, and otherwise as *a posteriori*. However, it was argued above that if the process is inferential at all, the relevant inferences are themselves of just the kind for which past experience plays a role that is neither purely enabling nor strictly evidential, so the inferential picture does not resolve the issue.

Suppose that we classify my knowledge of (25) in the envisaged circumstances as *a priori*, because sense experience plays no strictly evidential role; perhaps we insist on counting the role of such experience in knowledge of (25) as enabling. Then the danger is that far too much will count as *a priori*. Long-forgotten experience can mold my judgment in many ways without playing a direct evidential role, for example by calibrating my skilful application of concepts and conditioning me into patterns of expectation which are called on in my assessment of ordinary counterfactual conditionals. How we know (25) may turn out to be quite similar to how many of us know (26):

(26) If two marks had been nine inches apart, they would have been further apart than the front and back legs of an ant.

Sense experience need play no direct evidential role in knowledge of (26). One can know (26) without remembering any occasion on which one perceived an ant, and without having received any testimony about the size of ants. The ability to imagine accurately what an ant would look like next to two marks nine inches apart suffices. Doubtless (25) is necessary and (26) contingent. But that metaphysical difference does not imply any epistemological difference between how we know (25) and how we know (26). It does not justify the claim that (25) is known *a priori* and (26) *a posteriori*. Yet (26) is not usually supposed to be known or even knowable *a priori*.

Suppose, on the other hand, that we classify my knowledge of (25) as *a posteriori*, because experience plays more than a purely enabling

<sup>12</sup> Yablo (2002) has a related discussion of the concept *oval*.

role; perhaps we insist on counting the role of sense experience in knowledge of (26) as evidential. Then the danger is that the same verdict will apply to many philosophically significant modal judgments too. The assumption that they are known or even knowable *a priori* will be undercut. Of course, Kripke has argued strongly for a category of necessary truths knowable only *a posteriori*, such as “Gold is the element with atomic number 79”; “It is necessary that gold is the element with atomic number 79” would then be knowable only *a posteriori* too. The present suggestion is intended far more widely than that. For example:

- (27) It is necessary that whoever knows something believes it.
- (28) If Mary knew that it was raining, she would believe that it was raining.
- (29) Whoever knew something believed it.

Although (28) is not general and (29) is not modal, our way of knowing them is similar to our way of knowing (27); we do not learn (28) by analysis of Mary’s individual psychology or (29) by enumerative induction. Knowledge of truths such as (27)–(29) is usually regarded as *a priori*, even by those who accept the category of the necessary *a posteriori*. The experiences through which we learned to distinguish in practice between belief and non-belief and between knowledge and ignorance play no strictly evidential role in our knowledge of (27)–(29). Nevertheless, their role may be more than purely enabling. Many philosophers, native speakers of English, have denied (27) (Shope (1983: 171–92) has a critical survey). They are not usually or plausibly accused of failing to understand the words “know” and “believe.” Why should not subtle differences between two courses of experience, each of which sufficed for coming to understand “know” and “believe,” make for differences in how test cases are processed, just large enough to tip honest judgments in opposite directions? Whether knowledge of (27)–(29) is available to one may thus be highly sensitive to personal circumstances. Such individual differences in the skill with which concepts are applied depend constitutively, not just causally, on past experience, for the skillfulness of a performance depends constitutively on its causal origins.

In a similar way, past experience of spatial and temporal properties may play a role in skilful mathematical “intuition” that is not directly evidential but far exceeds what is needed to acquire the relevant

mathematical concepts. The role may be more than heuristic, concerning the context of justification as well as the context of discovery. Even the combinatorial skills required for competent assessment of standard set-theoretic axioms may involve offline applications of perceptual and motor skills, whose capacity to generate knowledge constitutively depends on their honing through past experience that plays no evidential role in the assessment of the axioms.

If the preceding picture is on the right lines, should we conclude that modal knowledge is *a posteriori*? Not if that suggests that (27)–(29) are inductive or abductive conclusions from perceptual data. In such cases, the question “*A priori* or *a posteriori*?” is too crude to be of much epistemological use. The point is not that we cannot draw a line somewhere with traditional paradigms of the *a priori* on one side and traditional paradigms of the *a posteriori* on the other. Surely we can; the point is that doing so yields little insight. The distinction is handy enough for a rough initial description of epistemic phenomena; it is out of place in a deeper theoretical analysis, because it obscures more significant epistemic patterns. We may acknowledge an extensive category of *armchair knowledge*, in the sense of knowledge in which experience plays no strictly evidential role, while remembering that such knowledge may not fit the stereotype of the *a priori*, because the contribution of experience was far more than enabling. For example, it should be no surprise if we turn out to have armchair knowledge of truths about the external environment.<sup>13</sup>

## 6

It is time to consider objections to the preceding account.

*Objection:* Knowledge of counterfactuals cannot explain modal knowledge, because the former depends on the latter. More specifically, in developing a counterfactual supposition, we make free use

<sup>13</sup> This problem for the *a priori/a posteriori* distinction undermines arguments for the incompatibility of semantic externalism with our privileged access to our own mental states that appeal to the supposed absurdity of *a priori* knowledge of contingent features of the external environment (McKinsey 1991). It also renders problematic attempts to explain the first dimension of two-dimensional semantics in terms of *a priori* knowability, as in Chalmers (2006). Substituting talk of rational reflection for talk of the *a priori* does not help, since it raises parallel questions.

of what we take to be necessary truths, but not of what we take to be contingent truths. Thus we rely on a prior or at least independent stock of modal knowledge or belief. The principle NECESSITY above illustrates how we do this.

*Reply:* Once we take something to be a necessary truth, of course we can use it in developing further counterfactual suppositions. But that does nothing to show that we have any special cognitive capacity to handle modality independent of our general cognitive capacity to handle counterfactual conditionals. If we start only with the latter, just as envisaged above, it will generate knowledge of various modal truths, which can in turn be used to develop further counterfactual suppositions, in a recursive process. For example, we need not judge that it is metaphysically necessary that gold is the element with atomic number 79 *before* invoking the proposition that gold is the element with atomic number 79 in the development of a counterfactual supposition. Rather, projecting constitutive matters such as atomic numbers into counterfactual suppositions is part of our general way of assessing counterfactuals. The judgment of metaphysical necessity originates as the output of a procedure of that kind; it is not an independently generated input.

What if our general cognitive capacity to handle counterfactuals has as a separate constituent a special cognitive capacity to handle metaphysical modality? Consider the cognitive resources sketched in Section 3 for the evaluation of counterfactual conditionals: most distinctively, imaginative simulation; less distinctively, reasoning, memory, testimony, perception. The question is whether they require supplementation by an additional capacity for the evaluation of counterfactuals of the special form  $A \Box \rightarrow \perp$ . They do not. Although we often cannot *perceptually* imagine the truth of  $A$ , not all imagining is perceptual imagining. “Imagine that there is a barber who shaves all and only those who do not shave themselves” is not radically different from the instruction “Suppose that there is a barber who shaves all and only those who do not shave themselves.” In imaginatively and inferentially developing a counterfactual supposition, one may or may not run into a contradiction. Of course, we often find claims of metaphysical possibility or necessity hard to evaluate. But that is not the point. There is no evidence whatsoever that we are *better* at evaluating claims of metaphysical modality than we would be if we had just the sorts of cognitive capacity listed above for

evaluating counterfactual conditionals, with no additional separate capacity for evaluating claims of metaphysical modality. Therefore the postulation of such an additional capacity is unwarranted.

*Objection:* The account associates metaphysical modality with counterfactual conditionals of a very peculiar kind: in the case of (17) and (18), those with an explicit contradiction as their consequent. Why should a capacity to handle ordinary counterfactuals confer a capacity to handle such peculiar ones too?

*Reply:* That is like asking why a capacity to handle inferences between complex empirical sentences should confer a capacity to handle inferences involving logical truths and falsehoods too. There is no easy way to have the former without the latter. More specifically, developing a counterfactual supposition includes reasoning from it, and we cannot always tell in advance when such reasoning will yield a contradiction (there are surprises in logic). The undecidability of logical truth for first-order logic implies that there is no total mechanical test for the consistency even of first-order sentences. Thus the inconsistent ones cannot be sieved out in advance (consider “In the next village there is a barber who shaves all and only those in that village who do not shave themselves”). Consequently, a general capacity to develop counterfactual suppositions must confer in particular the capacity to develop those which subsequently turn out inconsistent. Although the capacity may not be of uniform reliability, as already noted, the variation is primarily with the *antecedent* of the counterfactual (the supposition under development), not with its consequent (which is what is exceptional in (17) and (18)). In deductive inference, our reasoning to contradictions (as in proof by *reductio ad absurdum*) is not strikingly more or less reliable than the rest of our deductive reasoning. We can reach many conclusions about metaphysical modality without overstressing our imaginative resources. For instance, whenever we can deny a counterfactual  $A \Box \rightarrow B$ , we can assert  $\Diamond A$ , because  $A \Box \rightarrow \perp$  entails  $A \Box \rightarrow B$ . Again, the argument in Section 4 for a version of the necessity of identity employed only straightforward reasoning in the logic of counterfactuals. It is not an objection to the present account that our use of the imagination in evaluating counterfactuals may be unreliable for some with far-out antecedents.

*Objection:* The assumption about vacuous truth on which the account relies is wrong (Nolan 1997). For some counterpossibles

(counterfactuals with metaphysically impossible antecedents) are false, such as (30), uttered by someone who mistakenly believes that he answered “13” to “What is  $5 + 7$ ?”; in fact he answered “11”:

(30) If  $5 + 7$  were 13 I would have got that sum right.

Thus, contrary to (17),  $\Box A$  may be true while  $\neg A \Box \rightarrow \perp$  is false. In the argument for (17) in Section 3, the objectionable premise is NECESSITY. If some worlds are metaphysically impossible, and  $A$  is true at some of them but false at all metaphysically possible worlds, then every metaphysically possible  $A$  world is a  $B$  world, even if the closest  $A$  worlds are not  $B$  worlds.<sup>14</sup> Similar objections apply to the other purported equivalences (18)–(22).

*Reply:* Suppose that *all* counterpossibles are false. Then  $\Diamond A$  is equivalent to  $A \Box \rightarrow A$ , for the latter will still be true whenever  $A$  is possible; correspondingly,  $\Box A$  is equivalent to the dual  $\neg(\neg A \Box \rightarrow \neg A)$  and one can carry out the program of Section 3 using the new equivalences. But that is presumably not what the objector has in mind. Rather, the idea is that the truth-value of a counterpossible can depend on its consequent, so that (30) is false while (31) is true:

(31) If  $5 + 7$  were 13 I would have got that sum wrong.

However, such examples are quite unpersuasive.

First, they tend to fall apart when thought through. For example, if  $5 + 7$  were 13 then  $5 + 6$  would be 12, and so (by another eleven steps) 0 would be 1, so if the number of right answers I gave were 0, the number of right answers I gave would be 1. We prefer (31) to (30) because the argument for (31) is more obvious, but the argument for (30) is equally strong.

<sup>14</sup> Technically, NECESSITY fails on a semantics with similarity spheres for  $\Box \rightarrow$  that include some impossible worlds (inaccessible with respect to  $\Box$ ). Conversely, POSSIBILITY fails on a semantics with some possible worlds excluded from all similarity spheres (see Lewis (1986: 16) on universality). Inaccessible worlds seem not to threaten POSSIBILITY. For suppose that an  $A$  world  $w$  but no  $B$  world is accessible from a world  $v$ . Then if  $A \Box \rightarrow B$  holds at  $v$  on the usual semantics, there is an  $A$  world  $x$  such that every  $A$  world as close as  $x$  is to  $v$  is a  $B$  world. It follows that  $w$  is not as close as  $x$  is to  $v$  and that  $x$  is inaccessible from  $v$ , which contradicts the plausible assumption that any accessible world is at least as close as any inaccessible world.

Second, there are general reasons to doubt the supposed intuitions on which such examples rely. We are used to working with possible antecedents, and given the possibility of A, the incompatibility of B and C normally implies that  $A \Box \rightarrow B$  and  $A \Box \rightarrow C$  cannot both be true. Thus by over-projecting from familiar cases we may take the uncontentious (31) to be incompatible with (30). The logically unsophisticated make analogous errors in quantificational reasoning. Given the evident truth of “Every golden mountain is a mountain,” they think that “Every golden mountain is a valley” is false, neglecting the case of vacuous truth. Since the logic and semantics of counterfactual conditionals is much less well understood, even the logically sophisticated may find similar errors tempting. Such errors may be compounded by a tendency to confuse negating a counterfactual conditional with negating its consequent, given the artificiality of the constructions needed to negate the whole conditional unambiguously (“it is not the case that if . . .”). Thus the truth of  $A \Box \rightarrow \neg B$  (with A impossible) may be mistaken for the truth of  $\neg(A \Box \rightarrow B)$  and therefore the falsity of  $A \Box \rightarrow B$ . If we must choose between (30) and (31), it is clear which we should choose; but the impression that we must choose is an illusion.

Some objectors try to bolster their case by giving examples of mathematicians reasoning from an impossible supposition A (“There are only finitely many prime numbers”) in order to reduce it to absurdity. Such arguments can be formulated using a counterfactual conditional, although they need not be. Certainly there will be points in the argument at which it is legitimate to assert  $A \Box \rightarrow C$  (in particular,  $A \Box \rightarrow A$ ) but illegitimate to assert  $A \Box \rightarrow \neg C$  (in particular,  $A \Box \rightarrow \neg A$ ). But of course that does not show that  $A \Box \rightarrow \neg A$  is false. At any point in a mathematical argument there are infinitely many truths that it is not legitimate to assert, because they have not yet been proved (Lewis (1986: 24–6) pragmatically explains away some purported examples of false counterfactuals with impossible antecedents). Similarly, this reply could just as well have begun “If all counterpossibles were false,  $\Diamond A$  would be equivalent to  $A \Box \rightarrow A$ .” Read “the antecedent” in such a way that it is impossible. Then it would have been equally true to say “If all counterpossibles were false,  $\Diamond A$  would not be equivalent to  $A \Box \rightarrow A$ .” But that would not have mattered, for only the former counterfactual is assertable in a context in which for dialectical purposes the possibility of

the antecedent is not excluded, and that is what the argument requires.

We may also wonder what logic of counterfactuals the objectors envisage. If they reject elementary principles of the pure logic of counterfactual conditionals, that is an unattractive feature of their position. If they accept all those principles, then they are committed to operators characterized as in (17) and (18) that exhibit all the logical behavior standardly expected of necessity and possibility. What is that modality, if not metaphysical modality?

A final problem for the objection is this. Here is a paradigm of the kind of counterpossible the objector regards as false:

(32) If Hesperus had not been Phosphorus, Phosphorus would not have been Phosphorus.

Since Hesperus is Phosphorus, it is metaphysically impossible that Hesperus is not Phosphorus, by the necessity of identity. Nevertheless, the objectors are likely to insist that in imaginatively developing the counterfactual supposition that Hesperus is not Phosphorus, we are committed to the explicit denial of no logical truth, as in the consequent of (32). According to them, if we do our best for the antecedent, we can develop it into a logically coherent though metaphysically impossible scenario: it will exclude “Phosphorus is not Phosphorus.” But they will presumably accept this trivial instance of reflexivity:

(33) If Hesperus had not been Phosphorus, Hesperus would not have been Phosphorus.

In general, however, coreferential proper names are intersubstitutable in counterfactual contexts. For example, the argument from (34) and (35) to (36) is unproblematically valid:

(34) If the rocket had continued on that course, it would have hit Hesperus.  
(35) Hesperus = Phosphorus.  
(36) If the rocket had continued on that course, it would have hit Phosphorus.

Similarly, the argument from (33) and (35) to (32) should be valid. But (33) and (35) are uncontroversially true. If the objector concedes that (32) is true after all, then there should be an explanation of the felt resistance to it, compatible with its truth, and we may reasonably expect that explanation to generalize to other purported examples of false counterpossibles. On the other hand, if the objector rejects (32), they must deny the validity of the argument from (33) and (35) to (32). Thus they are committed to the claim that counterfactual conditionals create opaque contexts for proper names (the same argument could be given for other singular terms, such as demonstratives). But that is highly implausible. (34) and (36) are materially equivalent because their antecedents and consequents concern the same objects, properties, and relations: it matters not that different names are used, because the counterfactuals are not about such representational features (if the substitution of coreferential names in propositional attitude ascriptions does not preserve truth value, the reason is that such ascriptions are about representational features). But then exactly the same applies to (32) and (33). Their antecedents and consequents too concern the same objects, properties, and relations. That the antecedent of (32) and (33) is in fact metaphysically impossible does not radically alter their subject matter. The transparency of the counterfactual conditional construction concerns its general semantic structure, not the specific content of the antecedent.

Under scrutiny, the case for false counterpossibles looks feeble. The logic of quantifiers was confused and retarded for centuries by unwillingness to recognize vacuously true universal generalizations; we should not allow the logic of counterfactuals to be similarly confused by unwillingness to recognize vacuously true counterpossibles.<sup>15</sup>

*Objection:* Counterfactuals are desperately vague and context-sensitive; equivalences such as (17) and (18) will infect  $\Box$  and  $\Diamond$ , interpreted as metaphysical modalities, with all that vagueness and context-sensitivity.

*Reply:* Infection is not automatic. For instance, within a Lewis-Stalnaker framework, different readings or sharpenings of  $\Box \rightarrow$  may

<sup>15</sup> For an account of metaphysical modality in terms of counterfactuals that does admit false counterpossibles see Kment (2006). See also Lange (2005).

differ on the similarity ordering of worlds while still agreeing on what worlds there are, so that the differences cancel out in the right-hand sides of (17) and (18). Whether a given supposition counterfactually implies a contradiction may be unclear to us; that does not imply that there is no right answer.

On some dynamic accounts, the semantics of counterfactuals involves a more systematic interaction with context, because one normal effect of the antecedent is to update the context to one in which the horizon of contextually relevant worlds includes some in which the antecedent is true; the truth of the sentence is then equivalent to the truth of the consequent in *all* the relevant worlds in the updated context (von Fintel 2001). The present account can be adapted to such an account, if it is allowed that updating can fail to provide a world in which the antecedent is true when there is no such world, for then the counterfactual is vacuously true: its consequent is true in every relevant world in which its antecedent is true. Even if, less plausibly, the counterfactual is “undefined” in such cases (a view with awkward consequences for many informal mathematical proofs by *reductio ad absurdum* involving counterfactuals), metaphysical impossibility and the other modalities can still be recovered from the counterfactual, since “◊A” will be equivalent to “It is defined whether (A  $\Box\rightarrow$  A).”

*Objection:* It has been argued that counterfactual conditionals lack truth-values (Edgington 2003, Bennett 2003: 252–6). If so, the assimilation of claims of metaphysical possibility and necessity to counterfactuals will deprive such claims of truth-values.

*Reply:* The issues are too complex to discuss properly here, but the readily intelligible occurrence of counterfactual conditionals embedded in the scope of other operators as in (23) and (24) is hard to make sense of without attributing truth-values to the embedded occurrences. Here is another example:

(37) Every field that would have been flooded if the dam had burst was ploughed.

(37) can itself be intelligibly embedded in more complex sentences in all the usual ways; for example, it can be negated or made the antecedent of another conditional. In order to understand how such embeddings work, we must assign truth conditions to (37); *ad hoc*

treatments of a few particular embeddings are not enough. For (37) to have truth conditions, “field that would have been flooded if the dam had burst” must have application-conditions. Thus there must be a distinction between the fields to which “would have been flooded if the dam had burst” applies and those to which it does not. But that is just to say that there must be a distinction between the values of “ $x$ ” for which “If the dam had burst,  $x$  would have been flooded” is true and those for which it is false. That it is somewhat obscure what the truth conditions of counterfactual conditionals are, and that we sometimes make conflicting judgments about them, hardly shows that they do not exist. The requirement that counterfactual conditionals have truth conditions is one way in which the preceding discussion has not been perfectly neutral on their semantics.

## 7

The counterfactual conditional is of course not the only construction in ordinary use that is closely related to metaphysical modality. Consider comments after a swiftly extinguished fire in an explosives factory:

- (38) There could have been a huge explosion.
- (39) There could easily have been a huge explosion.

The truth-value of both (38) (on a natural reading) and (39) depends on the location of the fire, the precautions in place, and so on. The mere metaphysical possibility of a huge explosion is insufficient to verify either (38) (so interpreted) or (39). The restricted nature of the possibility is explicit in (39) with the word “easily”; it is implicit in the context of (38).<sup>16</sup> To discover the truth-value of (38) or (39), we need background information. We may also need our imagination, in attempting to develop a feasible scenario in which there is a huge explosion. We use the same general cognitive faculties as we do in evaluating related counterfactual conditionals, such as (40):

<sup>16</sup> On easy possibility see Sainsbury (1997), Peacocke (1999: 310–28) and Williamson (2000a: 123–30). On the idea that natural language modals such as “can” and “must” advert to contextually restricted ranges of possibilities see Kratzer (1977).

(40) If the fire engine had arrived a minute later, there would have been a huge explosion.

Judgments of limited possibility such as (38) (interpreted as above) and (39) have a cognitive value for us similar to that of counterfactual conditionals such as (40). Both (38) and (39) entail (41), although not *vice versa*:

(41) It is metaphysically possible that there was a huge explosion.

This is another way in which our ordinary cognitive capacities enable us to recognize that something non-actual is nevertheless metaphysically possible. But we cannot reason from the negation of (37) or of (38) to the negation of (40).

Can metaphysical possibility be understood as the limiting case of such more restricted forms of possibility? Perhaps, but we would need some account of what demarcates the relevant forms of possibility from irrelevant ones, such as epistemic possibility. It also needs to be explained how, from the starting-point of ordinary thought, we manage to single out the limiting case, metaphysical modality. The advantage of counterfactual conditionals is that they allow us to single out the limiting case simply by putting a contradiction in the consequent; contradictions can be formed in any language with conjunction and negation. Anyway, the connections with restricted possibility and with counterfactual conditionals are not mutually exclusive, for they are not being interpreted as rival semantic analyses, but rather as different cases in which the cognitive mechanisms needed for one already provide for the other.

The epistemology of metaphysical modality requires no dedicated faculty of intuition. It is simply a special case of the epistemology of counterfactual thinking, a kind of thinking tightly integrated with our thinking about the spatio-temporal world. To deny that such thinking ever yields knowledge is to fall into an extravagant skepticism. Here as elsewhere, we can do philosophy on the basis of general cognitive capacities that are in no deep way peculiarly philosophical.

# 6

## Thought Experiments

---

### 1

Of all the armchair methods of philosophy, one of the most conspicuous is the thought experiment. Much of the philosophical community allows that a judicious act of the imagination can refute a previously well-supported theory. In natural science, one might expect, to imagine obtaining a negative outcome to a crucial experiment may be to imagine refuting the theory at issue, but imagining refuting a theory no more actually refutes it than imagining killing a tyrant actually kills him. Why should philosophy be any different? If the idea of a crucial experiment is too crude to describe the workings of real science, that merely reinforces skepticism about crucial thought experiments in philosophy.

Such an objection to thought experiments is facile, as their seminal role in physics immediately suggests, most famously in the work of Galileo and Einstein. Of course, philosophy-hating philosophers (a common breed) claim that philosophical thought experiments are profoundly unlike those in natural science, in ways which make the former bad and the latter good, but we should be suspicious of such claims of philosophical exceptionalism. We have already seen the imagination play a mundane but vital role in the evaluation of counterfactual conditionals, from the most ordinary empirical ones to those equivalent to statements about metaphysical modality. We shall see it play a corresponding role in thought experiments.

The canonical example in the literature on philosophical thought experiments is Edmund Gettier's use of them to refute the traditional analysis of knowledge as justified true belief (Gettier 1963). The background working hypothesis is that his thought experiments are

paradigmatic, in the sense that if any thought experiments can succeed in philosophy, his do: thus to determine whether Gettier's thought experiments succeed is in effect to determine whether there can be successful thought experiments in philosophy. Even if we do not afford them quite that status, they provide a convenient focus for discussion. Moreover, they demonstrate the cognitive weight analytic philosophers rest on thought experiments. Sociologically, the phenomenon is remarkable. Gettier had no previous publications and was unknown to most of the philosophical profession; he did not write as an established authority. For the theory he was attacking, a neat and at the time widely accepted analysis of the central concept of epistemology, he cited well-known books by two leading philosophers of his time (Ayer 1956, Chisholm 1957) and, more tentatively, Plato (*Theaetetus* 201, *Meno* 98).<sup>1</sup> His three-page article turns on two imaginary examples.<sup>2</sup> Yet his refutation of the justified true belief analysis was accepted almost overnight by the community of analytic epistemologists. His thought experiments were found intrinsically compelling.

This chapter analyzes the logical structure of Gettier-style thought experiments. The discussion can be generalized to many imaginary counterexamples that have been deployed against philosophical analyses and theories in ways more or less similar to Gettier's. Far more extensive investigation would be needed to warrant the claim that all philosophical thought experiments work in that way, but one must start somewhere. The main overall aim is to subsume the epistemology of thought experiments under the epistemology of counterfactual conditionals and metaphysical modality developed in the previous chapter, and thereby to reveal it as an application of quite ordinary ways of thinking, not as something peculiarly philosophical. A related subsidiary aim is to achieve a finer-grained understanding of the structure of the arguments that underlie thought experiments, both

<sup>1</sup> Shope (1983: 12–19) discusses whether Plato endorsed the justified true belief analysis of knowledge and argues that Kant did in the *Critique of Pure Reason* at A822, B850.

<sup>2</sup> Russell (1912) gave examples with a structure very similar to Gettier's, but used them only to draw the conclusion that “a true belief is not knowledge when it is deduced from a false belief” (in the chapter on “Knowledge, error and probable opinion”).

for its own sake and in order to test the overall account by developing it in detail.

## 2

We can extract from Gettier's paper an argument that makes no obvious appeal to thought experiments. According to the target analysis, a necessary and sufficient condition for knowing something is that it is true, one believes it and one is justified in believing it; for short, one has justified true belief.<sup>3</sup> Now in the sense of "justified" in which being justified in believing something is necessary for knowing it, Gettier argues, one can be justified in believing what is in fact false (the truth component of the justified true belief analysis is not redundant). But if one is justified in believing something, and correctly deduces from it something else, one is justified in believing the latter proposition on that basis (deduction is a way of transmitting justification from the premises to the conclusion of an argument). Since any truth is deductively entailed by various falsehoods, one can believe a truth on the basis of having correctly deduced it from a falsehood one is justified in believing, and thereby be justified in believing the deduced truth too; thus one has justified true belief in the latter. Nevertheless, one does not know, for one's belief in the truth, no matter how justified, is essentially based on a false lemma; one's conclusion cannot be epistemically better off than one's premises. Therefore, justified true belief is insufficient for knowledge.

One disadvantage of the abstract argument is that it rests on several very general claims for which we might find adequate support hard to provide. In particular, it assumes that a belief essentially based on a false belief does not constitute knowledge. Can we take that for granted? How do we know that a belief essentially based on a false belief never constitutes knowledge even in *recherché* cases? Fortunately, the universal generalization is more than Gettier needs in order to refute the target analysis. He needs only some particular instance in which the belief essentially based on a false lemma clearly

<sup>3</sup> In a sense, one can believe something and be justified in believing it without having a justified belief in it, because the available justification is not the reason for which one believes. What follows does not depend on this distinction.

fails to constitute knowledge, whether or not all other cases go the same way. As Gettier proceeds, the verdict that the subject lacks knowledge in the particular case has epistemic priority over the general diagnosis that a true belief essentially based on a false one never constitutes knowledge. In this account, the primary direction of support is abductive, from particular verdict to general principle (by inference to the best explanation), rather than deductive, from general principle to particular verdict (by universal instantiation). Gettier's own focus is on the particular verdicts, and that is how his counterexamples have usually been understood as working. In any event, his examples *can* be used in that way, and methodologically it is best to start with the simplest case, in which the particular verdict has priority. A similar point applies to Gettier's explicit assumption that justification is closed under deduction: what matters for his immediate purposes is just that the assumption clearly holds in his chosen cases, whether or not it holds in all more *recherché* ones. The need for examples is also implicit in Gettier's claim that one can be justified in believing a falsehood, for how could he adequately support the claim without appeal to examples? In effect, he provides a general recipe for developing any example of justified false belief into a counterexample to the justified true belief analysis.

Gettier's assumption that there can be justified false belief is not unquestionable, for any belief which does not constitute knowledge is *ipso facto* defective, and so in some sense not fully justified, even if it is fully excusable. That objection clearly does not invoke a standard of justification on which it is unnecessary for knowledge, nor does it give any succor to skepticism. However, it does invoke a concept of justification which is not prior to the concept of knowledge, and so risks making the analysis of knowledge as justified true belief circular (Williamson 2000a: 184–5, Sutton 2007). The analysis of knowledge as justified true belief loses much of its intended explanatory power if justification has to be understood by reference to knowledge. It is dialectically legitimate for critics of the analysis to work, as Gettier does, with the view of justification on which its proponents rely. On such a view, my justification for believing that I have hands is equally good whether I am an ordinary human with hands or a brain in a vat which merely seems to itself to be an ordinary human with hands: since my belief is justified in the former case, on pain of skepticism, it is equally justified in the latter case, when

it is false. In what follows, we assume a sense of “justified” in which one can be justified in believing falsehoods.

Gettier presents his specific counterexamples to the target analysis through short fictional narratives, in the present tense indicative, with fictional uses of proper names (“Smith” and “Jones”), all introduced by “suppose that.” Beyond their conformity to the abstract pattern just explained, their details do not concern us. Let us construct another example to the same pattern. A clever bookseller fakes evidence which appears to show conclusively that a particular book once belonged to Virginia Woolf; convinced, Orlando pays a considerable sum for the book. He has a justified false belief that this book of his once belonged to Virginia Woolf. On that basis alone, he forms the existential belief that he owns a book which once belonged to Virginia Woolf. The latter belief is in fact true, because another of his books in fact once belonged to her, although he does not associate that one with her in any way. Thus Orlando has a justified true belief that he owns a book which once belonged to Virginia Woolf, but he does not know that he owns a book which once belonged to Virginia Woolf. What we need to understand is how such fictional narratives can present counterexamples to philosophical analyses.

On Gettier’s account, the target analysis is a claim of necessary and sufficient conditions for knowing. Let us formalize this as the claim that, necessarily, for any subject  $x$  and proposition  $p$ ,  $x$  knows  $p$  if and only if  $x$  has a justified true belief in  $p$ .<sup>4</sup> Symbolically:

$$(1) \quad \Box \forall x \forall p (K(x, p) \equiv JTB(x, p))$$

This does not say that knowledge is identical with justified true belief, nor does it entail that the word “knowledge” is synonymous with the phrase “justified true belief” or that the concept *knowledge* is identical with the concept *justified true belief*. But if any of those further claims is true, so too is (1). Thus a refutation of (1) automatically refutes each of those further claims too, although not conversely.

For present purposes, in formalizing Gettier’s argument against (1), we can ignore most of the structure specific to his cases, and

<sup>4</sup> The assumption that propositions are the objects of knowledge is convenient, but inessential to the underlying argument.

concentrate on the logical structure they share with most other imaginary counterexamples to philosophical analyses. Suppose that we fix on a particular Gettier-style story (the one about Orlando would do), henceforth “the Gettier case,” told in neutral terms, without prejudice to the target analysis. For instance, it is not explicitly part of the story that Orlando does not know that he owns a book which once belonged to Virginia Woolf. Since the story contains fictional singular terms, such as “Orlando” and “this book,” it is arguably just a pretense that its constituent sentences express propositions. However, we can treat such fictional singular terms as picturesque substitutes for variables. Replacing them by variables, we can represent the Gettier-style story by an open sentence  $GC(x, p)$ , where the variables “ $x$ ” and “ $p$ ” occupy the positions for, respectively, the believer and the content of the justified true belief. Although one could attempt an analysis of thought experiments that took their fictional aspect more seriously, their relevance to fictional claims such as (1) is most easily understood in this more literal-minded way.

If the Gettier case were impossible, it would pose no obvious threat to the claim of necessity (1). We therefore make the putative possibility of the case explicit:

$$(2) \quad \Diamond \exists x \exists p \ GC(x, p)$$

Someone could stand in the relation described in the Gettier story to some proposition. In order to complete the argument against (1), we need the verdict that the subject in the Gettier case has justified true belief without knowledge. To a first approximation, we can formalize that as the claim that, necessarily, anyone who stands in the Gettier relation to a proposition has justified true belief in that proposition without knowledge:

$$(3) \quad \Box \forall x \forall p \ (GC(x, p) \rightarrow (JTB(x, p) \ \& \ \neg K(x, p)))$$

By elementary modal reasoning, a necessary consequence of something possible is itself possible. Therefore, as a logical consequence of (2) and (3), someone could have justified true belief in a proposition without knowledge:

$$(4) \quad \Diamond \exists x \exists p \ (JTB(x, p) \ \& \ \neg K(x, p))$$

But (4) is straightforwardly inconsistent with (1), in particular with its right-to-left direction. Justified true belief is insufficient for knowledge. Consequently, (2) and (3) suffice as premises for a deductive argument against the target analysis.

This objection to (1) relies essentially on its modal content. If (1) were replaced by a non-modal universally quantified biconditional, thought experiments would not refute it, for an imaginary case in which two things fail to coincide is quite compatible with their coincidence over all actual cases. The function of the thought experiment is to show that a certain case could arise, and that if it did, the two things would come apart, from which it follows that the two things could come apart. That refutes the modal claim that they could not come apart, but not the non-modal claim that they never in fact come apart.

That (3) is the best representation of the verdict on the Gettier case is doubtful. In philosophy, examples can almost never be described in complete detail. An extensive background must be taken for granted; it cannot all be explicitly stipulated. Although many of the missing details are irrelevant to whatever philosophical issues are in play, not all of them are. This applies not just to highly schematic descriptions of examples, such as the initial abstract Gettier schema, but even to the much richer stories Gettier and other philosophers like to tell. For example, in the Gettier case, if the subject's inference to the true belief  $p$  from the false belief  $q$  bizarrely happens to trigger awkward memories or apparent memories that cast doubt on  $q$ , the effect may be to lose justification for  $q$  rather than to gain it for  $p$ . Without specifically addressing the question, we do not envisage the Gettier case like that. Nor do we worry about whether our verdicts would hold even if mad scientists were interfering with the subject's brain processes in various ways; those possibilities do not normally occur to us when we assess Gettier examples. Similarly, when moral philosophers assess imaginary examples, one can almost always fill out the case with unintended but morally relevant additions that would reverse the verdict. Any humanly compiled list of such interfering factors is likely to be incomplete.

Instead of asking whether justified true belief without knowledge is a necessary consequence of the Gettier case, one might more naturally ask whether, if there *were* an instance of the Gettier case, it *would* be an instance of justified true belief without knowledge. The

verdict that it would constitutes a counterfactual conditional, which is much weaker than the strict conditional (3).<sup>5</sup> In very rough terms, it requires justified true belief without knowledge only in the closest realizations of the Gettier case, not in all possible realizations. By using the counterfactual conditional, we in effect leave the world to fill in the details of the story, rather than trying to do it all ourselves. For present purposes the counterfactual can be symbolized thus (its formalization will be discussed in detail later):

$$(3^*) \quad \exists x \exists p \, GC(x, p) \square\rightarrow \forall x \forall p \, (GC(x, p) \rightarrow (JTB(x, p) \ \& \ \neg K(x, p)))$$

The counterfactual conditional in (3<sup>\*</sup>) takes widest possible scope. If there were an instance of the Gettier case, it would be an instance of justified true belief without knowledge. For the time being, let us simply assume that (3<sup>\*</sup>) correctly formalizes Gettier's major premise. That assumption will be evaluated in later sections.

Let us reconstruct the logic of the argument against (1). Informally, why do (2) and (3<sup>\*</sup>) entail (4)? Given (2), (3<sup>\*</sup>) cannot hold vacuously. Thus, given (2) and (3<sup>\*</sup>), (3<sup>\*</sup>) holds non-vacuously. Therefore its antecedent and consequent hold together in some possible world. That must be a possible world in which someone has justified true belief without knowledge. Thus (4) is true, so (1) is false.

We can make the reasoning rigorous without reliance on possible worlds. First, consider the logical relations between the non-modal constituents of the argument. Let A be  $\exists x \exists p \, GC(x, p)$  ("Someone stands in the Gettier relation to something"), B be  $\forall x \forall p \, (GC(x, p) \rightarrow (JTB(x, p) \ \& \ \neg K(x, p)))$  ("Whoever stands in the Gettier relation to something has justified true belief in it without knowledge") and C be  $\exists x \exists p \, (JTB(x, p) \ \& \ \neg K(x, p))$  ("Someone has justified true belief in something without knowledge"). Thus (2) is  $\Diamond A$ , (3<sup>\*</sup>) is  $A \square\rightarrow B$  and (4) is  $\Diamond C$ . Obviously, C is a logical conse-

<sup>5</sup> Similarly, in describing one of his famous examples to motivate the causal theory of perception, Grice writes "if, unknown to me, there were a mirror interposed between myself and the pillar, it would certainly be incorrect to say that I saw the first pillar, and correct to say that I saw the second" (1961, section 5); the counterfactual conditional here reads completely naturally (although one might object to his dressing up a fact about perception as, in his words, a "linguistic fact"). Sorensen (1992) formalizes the arguments underlying thought experiments using counterfactual conditionals; for discussion of his proposal see Häggqvist (1996: 92–103).

quence of  $A$  and  $B$ : in symbols,  $A, B \vDash C$ . By the principle that the counterfactual consequences of a given supposition are closed under logical consequence (CLOSURE), we therefore have  $A \Box \rightarrow A$ ,  $A \Box \rightarrow B \vDash A \Box \rightarrow C$ .<sup>6</sup> Since everything counterfactually implies itself (REFLEXIVITY), the first premise is a logical truth:  $\vDash A \Box \rightarrow A$ . Thus we can simplify to  $A \Box \rightarrow B \vDash A \Box \rightarrow C$ . By the principle POSSIBILITY from the previous chapter, a counterfactual consequence of a possibility is itself a possibility, which yields  $\Diamond A, A \Box \rightarrow C \vDash \Diamond C$ . Combining these two results gives  $\Diamond A, A \Box \rightarrow B \vDash \Diamond C$ , in other words, (2), (3\*)  $\vDash$  (4), as required. Thus weakening the major premise from a strict to a counterfactual implication leaves the validity of the argument intact. The extra strength of strict implication was an unnecessary commitment.

This account of the use of imaginary counterexamples in refuting philosophical analyses extends far beyond Gettier cases. It also generalizes to their use in refuting philosophical claims of necessity which lack the form of an analysis, such as one-way strict implications.

*Preview:* Section 3 makes some observations about the epistemology of the argument just analyzed. Section 4 assesses (3\*) as a formalization of the counterfactual (Appendix 2 considers another alternative). Section 5 asks whether the right counterfactual was selected for formalization. The final section considers whether Gettier's argument concerns counterfactual possibility at all. The foregoing account survives all these tests, at least as an adequate approximation.

### 3

On our account, a thought experiment such as Gettier's embodies a straightforward valid modal argument for a modal conclusion. The role of the imagination is in verifying the premises.<sup>7</sup>

<sup>6</sup> As noted in the previous chapter, CLOSURE cannot be applied to cases when the original argument preserves truth at the actual world of every model but not at counterfactual worlds. Since  $C$  is an ordinary first-order logical consequence of  $A$  and  $B$ , this problem does not arise here.

<sup>7</sup> There is a debate as to whether thought experiments in science reduce to arguments (Norton 1991, 2004) or contain an irreducible imaginative element (Gendler

The major premise (3\*) is a counterfactual conditional; the imagination is used in verifying it just as it is used in verifying many everyday counterfactuals, such as “If the bush had not been there, the rock would have landed in the lake.” There is nothing peculiarly philosophical about the way in which the counterfactual is assessed. The antecedent and consequent express empirical conditions. Is the connection between them endorsed on distinctively “conceptual” grounds? The epistemological idea of conceptual connections turned out in earlier chapters to be a myth. Here, two points are enough. First, if what warranted the counterfactual conditional (3\*) was that its antecedent conceptually entailed its conclusion, then that would also warrant the strict implication (3); but we have seen that the strict implication is not warranted. Second, native English speakers sometimes dispute the Gettier verdict, and so by implication reject the counterfactual. In doing so, they show poor epistemological judgment but not linguistic incompetence: they are not usually accused of failing to understand the relevant words of English; it would be inappropriate to send them off to language school for retraining. Some of them have had no exposure to philosophy; others are professional epistemologists.<sup>8</sup> We assent to (3\*) on the basis of an offline application of our ability to classify people around us as knowing various truths or as ignorant of them, and as having or as lacking other epistemologically relevant properties. That classificatory ability goes far beyond mere linguistic understanding of “know” and other words.

The minor premise (2) is a claim of possibility. For standard Gettier cases it is quite uncontroversial. They constitute mundane practical possibilities; nobody doubts that they could arise: (2) is not where the philosophical action is. What skeptics about Gettier’s thought experiments doubt is not (2) but (3\*). They call into question “the Gettier intuition,” that the case is one of justified true belief without knowledge: it corresponds to (3\*), not (2), for the English original of (2) does not even contain “know” or cognate terms. In

---

1998, 2004). The present account of thought experiments in philosophy goes some way towards reconciling the two sides: thought experiments do constitute arguments, but the imagination plays an irreducible role in warranting the premises.

<sup>8</sup> Shope (1983: 26–33) discusses some attempts by professional epistemologists to argue that the Gettier problem is not genuine. See Weinberg, Stich, and Nichols (2001) for lay denials.

any case, the previous chapter showed how the ordinary epistemology of counterfactual conditionals applies to possibility claims such as (2).

For other philosophical thought experiments, the possibility premise corresponding to (2) may be far more contentious: a bizarre science fiction possibility, perhaps involving a brain swap or even a disembodied mind. Whether the possibility premise is warranted depends on the details of the case, but there is no reason in principle why it should not be. In general, we have a trade-off between the uncontentiousness of the major premise and the uncontentiousness of the minor premise. The more we pack into the description of the case (such as  $GC(x, p)$ ), the more firmly we can secure the major premise, the desired verdict, but the less obvious we make the minor premise, the possibility claim. By packing less into the description, we can make the possibility claim more obvious, but risk loosening our grip on the desired verdict. However, such trade-offs are a commonplace of abstract argument; they do not mean that we cannot make both premises simultaneously plausible enough for our purposes.

Do we know the premises (2) and (3\*) *a priori*? Presumably, we do so if and only if we also know the conclusion (4) *a priori*, given that we believe it just on the basis of this logically valid deduction. However, in the previous chapter we saw reason to doubt the significance of the distinction between *a priori* and *a posteriori* knowledge. The considerations there apply to the present case too. We accept (2) and (3\*) on the basis of a capacity for applying epistemological concepts that goes far beyond what it takes to possess the concepts in the first place, since someone with a distorted epistemological outlook may reject (3\*), yet still possess the relevant concepts: they genuinely believe that the subject of the Gettier case would not have justified true belief without knowledge. Past experience contributed to the acquisition of those classificatory epistemological skills that go far beyond possession of the relevant concepts. That experience included sense experience. For example, we learn to recognize perceptually conditions of observation under which observers can gain perceptual knowledge of various features of their environment. Again, our skill in discriminating justification from its absence is developed in observation of other thinkers. In our acceptance of (3\*), sense experience is not confined to a purely enabling role, for example by providing

the opportunity to acquire those concepts or to encounter philosophical arguments about them. It is more directly implicated than that. It plays a positive role in helping to tip judgment one way rather than the other when one imagines the Gettier case instantiated as such that the subject's inference extends justification from the false premise to the true conclusion, rather than as such that the inference undermines justification for the premise. Which way one goes depends on what one finds normal or natural, which partly depends on the past course of one's sense experience. Thus knowledge of (3\*) does not conform to the usual stereotype of *a priori* knowledge. Typically, however, past experience plays no strictly evidential role in knowledge of (3\*): for example, we need not invoke past instances of lack of knowledge as inductive evidence for lack of knowledge in the Gettier case. The experience of performing the thought experiment itself is not sense experience as usually understood. Thus knowledge of (3\*) fails equally to conform to the usual stereotype of *a posteriori* knowledge. Although we might try to resolve the issue by stipulation, doing so would yield little insight into the nature of knowledge such as we have of (3\*). To gain such insight, we must focus on the ways in which that knowledge differs both from the stereotype of *a priori* knowledge and from the stereotype of *a posteriori* knowledge.

One manifestation of the influence of past experience on epistemological judgments may be cross-cultural variation in verdicts on thought experiments, including the Gettier case.<sup>9</sup> Such variation, if it occurs, may result from cross-cultural variation in the meaning of "know" or other epistemological terms, but it need not. It may occur between sub-communities of English speakers who all use the words as part of a single common vocabulary, but disagree in their applications of them, just as different communities may disagree in their applications of the word "justice" while still using it with a single shared meaning. Cross-cultural disagreement over the theory of evolution is compatible with a common meaning of the word "evolution" between the cultures. In the present cases, the variation between individuals within a single group is just as striking as the statistical

<sup>9</sup> For some evidence see Weinberg, Stich, and Nichols (2001), critically discussed by Sosa (2005). The rationale for the use of thought experiments in philosophy which Weinberg, Stich and Nichols attack is very different from that defended in this book.

variation between groups: the data do not suggest a clash of monolithic cultures, but rather some variation in the proportion of the population who respond in a given way.

Much of the evidence for cross-cultural variation in judgments on thought experiments concerns verdicts by people without philosophical training. Yet philosophy students have to learn how to apply general concepts to specific examples with careful attention to the relevant subtleties, just as law students have to learn how to analyze hypothetical cases. Levels of disagreement over thought experiments seem to be significantly lower among fully trained philosophers than among novices. That is another manifestation of the influence of past experience on epistemological judgments about thought experiments.

We should not regard philosophical training as an illegitimate contamination of the data, any more than training natural scientists how to perform experiments properly is a contamination of their data. Although the philosophically innocent may be free of various forms of theoretical bias, just as the scientifically innocent are, that is not enough to confer special authority on innocent judgment, given its characteristic sloppiness. Training in any intellectual discipline whatsoever has some tendency to instill unquestioning conformity to current basic assumptions in that discipline, and a consequent slowness to recognize errors in those assumptions. That is inevitable, for no progress is made when everything is put simultaneously into question. Fully trained practitioners can still obtain experimental results that undermine currently accepted theories. That can happen with philosophical thought experiments too, as the example of Gettier shows.<sup>10</sup>

The residual levels of disagreement in judgments between trained philosophers do not warrant wholesale skepticism about the method of thought experiments. Naturally, philosophical debates focus on points of disagreement, not on points of agreement. Most intellectual disciplines have learned to live with significant levels of disagreement between trained practitioners, concerning both theory and observation: philosophy is not as exceptional in this respect as some pretend.

<sup>10</sup> Contrast Goldman (2005), discussed in Kornblith (2007). Goldman's interest is in the analysis of "pretheoretic concepts," but theoretical innocence often causes people to misapply their own concepts.

Notoriously, eye-witnesses often disagree fundamentally in their descriptions of recent events, but it would be foolish to conclude that perception is not a source of knowledge, or to dismiss all eye-witness reports. To ignore the evidence of thought experiments would be a mistake of the same kind, if not of the same degree. Disagreement can provide a reason to be somewhat more cautious than we might otherwise have been, in our handling both of eye-witness reports and of thought experiments; such caution is commonplace in philosophy. There is no need to be panicked into more extreme reactions.

This account has emphasized the epistemological continuity between verdicts on philosophical thought experiments and other judgments. That emphasis is supported by cases in which observations of real life do the same epistemological work as philosophical thought experiments. For instance, not all Gettier counterexamples are imaginary: sometimes a stopped watch really does show the right time. To make the point vivid, I have occasionally created Gettier cases for lecture audiences. For example, I have begun a lecture by apologizing for not giving a power-point presentation; I explained that the only time I gave a power-point presentation it was a complete disaster. Since my listeners had no reason to distrust me on a claim so much to my discredit, they acquired through my testimony the justified belief that the only time I gave a power-point presentation it was a complete disaster. They competently deduced that I had never given a successful power-point presentation. Thus they acquired the justified belief that I had never given a successful power-point presentation. That belief was true, but the reason was that I had never given a power-point presentation at all (and still do not intend to). My assertion that the only time I had given a power-point presentation it was a complete disaster was a bare-faced lie.<sup>11</sup> Thus they were basing their justified true belief that I had never given a successful power-point presentation on their justified false belief that the only time I had given a power-point presentation it was a complete disaster. Consequently, they did not know that I had never given a successful power-point presentation. The original audience encountered the case by living through it, others do so by reading my testimony, which is more similar to encountering a case by reading a fictional narrative. Either way, this actual Gettier case is a counterexample to

<sup>11</sup> Someone commented “You can’t believe the first thing he says.”

the non-modal principle that knowledge coincides with justified true belief in all actual cases; since actuality entails possibility ( $A \models \Diamond A$ ), it is also a counterexample to the modal principle (1) that knowledge coincides with justified true belief in all possible cases.

What is striking about real life Gettier cases is how little difference they make. They are not markedly more or less effective as counter-examples to the target analysis than imaginary Gettier cases are. Those who found the imaginary counterexamples convincing find the real life ones more or less equally convincing. Unless one is a skeptic about the external world, the reliance on empirical methods is no reason for serious doubt.<sup>12</sup> Conversely, those who were suspicious of the imaginary counterexamples are more or less equally suspicious of the real life ones.

It might be replied that the process of classifying a real life instance of the Gettier case as an instance of justified true belief without knowledge involves a modal judgment, because it can be factorized into a deduction from the non-modal premise that this is an instance of the Gettier case and the modal premise that if something were an instance of the Gettier case it would be an instance of justified true belief without knowledge. However, such factorization is deeply problematic. Note first that the modal element in it is quite gratuitous; the deduction works just as well with the non-modal second premise that every (actual) instance of the Gettier case is an instance of justified true belief without knowledge. Furthermore, we have no good reason to insist on factorization here but not for utterly ordinary ascriptions of epistemological predicates, as when someone says that John does not know that the meeting has been cancelled. Nor have we any good reason to insist on it for ascriptions of epistemological predicates and not for ascriptions of other empirical predicates. But if factorization is ubiquitous, an infinite regress occurs. The process of classifying this as an instance of the Gettier case is itself factorized into a deduction from the non-modal premise that this is an instance of F and the modal premise that if something were an instance of F it would be an instance of the Gettier case. The process of classifying this as an instance of F would in turn be factorized into

<sup>12</sup> In principle, someone could react to a real life Gettier case by judging it to be possible without judging it to be actual, and reject (1) on the former basis alone. It is implausible that most people take that unnatural route.

a deduction from the non-modal premise that this is an instance of E and the modal premise that if something were an instance of E it would be an instance of F, and so on. Plainly, no such infinite regress of inferences occurs in us. At some point, we simply apply our concepts to what confronts us, without relying on an inference from further premises. Why should that not happen with the original epistemological classification of the real life instance of the Gettier case? No doubt epistemological facts supervene on non-epistemological facts (so that the non-epistemological facts in a suitable instance of the Gettier case determine that it is an instance of justified true belief without knowledge), but of course that does not entail that our epistemological beliefs are derived from non-epistemological beliefs. Our epistemological beliefs are certainly not inferred from our beliefs about microphysics, even if epistemological facts supervene on microphysical facts. Why should our epistemological beliefs be inferred from some other putative supervenience base? Most people have scarcely any idea how to formulate even approximately sufficient conditions in informative non-epistemological terms for epistemological conclusions. Even if they do happen to speculate along such lines, their speculations are far less secure epistemically than are their ordinary applications of epistemological concepts, so the latter do not depend on the former. The factorization hypothesis has little independent plausibility. Moreover, even if the factorization hypothesis were true, it would apply equally to non-philosophical applications of epistemological predicates in ordinary life and natural science, and so would indicate nothing distinctive about their applications in real-life instances of Gettier cases.

Removing the tricky apparatus of thought experiments and modal judgments does not reassure those who doubted that the subjects of Gettier's original examples lacked knowledge: whatever their rhetoric, their doubts did not really concern the method of thought experiments. Rather, they concerned the reliability of our epistemological judgments, whether modal or non-modal, in particular of our applications of the concepts of knowledge and justification.<sup>13</sup> The switch

<sup>13</sup> *Objection:* Nozick (1981: 172–96) analyzes knowledge in counterfactual terms; on his view, any judgment about knowledge implicitly involves judgments concerning counterfactual conditionals. *Reply:* First, the objection does not fully generalize, since it depends on a specific analysis of knowledge. Second, Nozick's analysis does not make philosophers' ascriptions of knowledge or of its absence any more modal than

from an “*a priori*” to an “*a posteriori*” method here makes very little practical difference. We manifest recognition of this underlying cognitive similarity when we refuse to treat real life and fictional Gettier cases as mutually independent evidence against the justified true belief account of knowledge to a much greater extent than we treat two fictional Gettier cases as mutually independent evidence.

## 4

Let us now consider more carefully the fine structure of the major premises of the arguments which underlie philosophical thought experiments.

What is sometimes called “the Gettier intuition” has been expressed by a counterfactual conditional in English, roughly:

(5) If a thinker were Gettier-related to a proposition, he/she would have justified true belief in it without knowledge.<sup>14</sup>

This was in turn symbolized by the formula (3\*). Later, we will assess some alternative expressions of the Gettier intuition in English. For the time being, let us treat (5) as faithfully expressing the Gettier intuition, and ask whether (3\*) faithfully enough formalizes (5).

We might query (3\*) as a formalization of (5) on grounds of syntactic structure. Where (5) has the anaphoric pronouns “he/she” and “it,” (3\*) repeats the material  $GC(x, p)$  and applies universal quantification. In fact, (5) is a case of “donkey anaphora.” It is similar to (6):

(6) If a farmer owned a donkey, he would beat it.

---

non-philosophers’ ascriptions are. Third, skeptics about epistemological thought experiments typically make no appeal to counterfactual analyses of knowledge. After all, the way in which Nozick reaches his conclusions exemplifies the very methodology about which they are skeptical. Nor would they regard their skepticism as undermined by growing evidence that counterfactual analyses of knowledge are incorrect (Williamson 2000a: 147–63). Their skepticism is intended to get its grip irrespective of whether ascriptions of knowledge as such involve modal thinking.

<sup>14</sup> To be Gettier-related here is to be related as specified in the given Gettier scenario, not merely to be related as in some Gettier scenario or other.

This is just the “subjunctive” analogue of the classic indicative donkey sentence (7):

(7) If a farmer owns a donkey, he beats it.

The standard first-order formalization of (7) is (8):

(8)  $\forall x \forall y ((\text{Farmer}(x) \ \& \ \text{Donkey}(y) \ \& \ \text{Owns}(x, y)) \rightarrow \text{Beats}(x, y))$

The main challenge is to explain how (7) can have the truth conditions of (8) in terms of a compositional semantics for (7), given the mismatch in syntactic structure between (7) and (8).<sup>15</sup> For present purposes, however, what matters most is just getting the right truth conditions, up to logical equivalence. We might expect that if (7) has the same truth conditions as (8), then (6) will have the same truth conditions as the result of replacing the material conditional in (8) by a counterfactual conditional:

(9)  $\forall x \forall y ((\text{Farmer}(x) \ \& \ \text{Donkey}(y) \ \& \ \text{Owns}(x, y)) \Box \rightarrow \text{Beats}(x, y))$

The analogous formalization of (5) is not (3\*) but (10):

(10)  $\forall x \forall p (\text{GC}(x, p) \Box \rightarrow (\text{JTB}(x, p) \ \& \ \neg \text{K}(x, p)))$

In the indicative case, (8) is logically equivalent to the donkey analogue of (3\*):

(11)  $\exists x \exists y (\text{Farmer}(x) \ \& \ \text{Donkey}(y) \ \& \ \text{Owns}(x, y)) \rightarrow$   
 $\forall x \forall y ((\text{Farmer}(x) \ \& \ \text{Donkey}(y) \ \& \ \text{Owns}(x, y)) \rightarrow \text{Beats}(x, y))$

<sup>15</sup> Elbourne (2005) is a recent discussion of the topic with further references. Some will judge “If John had a dime, he would put it in the meter” true if in the relevant counterfactual circumstances John has two dimes and puts one in the meter. They may also have to judge “If John had a dime, he would put it in his piggybank” simultaneously true by parity of reasoning. There is no corresponding true reading of “If John had a dime, he would put it in the meter and put it in his piggybank.” All that is clearly true in the envisaged case is “If John had a dime, he would put one in the meter.”

For since (8) is the consequent of (11), (8) obviously entails (11), and conversely, if the antecedent of (11) is false then (8) is vacuously true, so (11) entails (8). However, the corresponding equivalence fails in the counterfactual case: (9) is not equivalent to (12).

$$(12) \quad \exists x \exists y (\text{Farmer}(x) \ \& \ \text{Donkey}(y) \ \& \ \text{Owns}(x, y)) \square \rightarrow \forall x \forall y ((\text{Farmer}(x) \ \& \ \text{Donkey}(y) \ \& \ \text{Owns}(x, y)) \rightarrow \text{Beats}(x, y))$$

For (12) is true and (9) false in the following circumstances. In the actual world (and, if you like, in all close ones) some farmer owns some donkey and every farmer who owns a donkey beats it. Farmer Giles could have owned this donkey, although he does not own it in the actual world (or in any close one). If he owned it, he would not beat it. Similarly, (10) is not equivalent to (3\*). For (3\*) is true and (10) false in the following circumstances. In the actual world (and, if you like, in all close worlds) someone is Gettier-related to some proposition and everyone who is Gettier-related to a proposition has justified true belief in it without knowledge. That woman could have been Gettier-related to that proposition, although she is not Gettier-related to it in the actual world (or in any close world). If she had been Gettier-related to it, she would have lacked justified belief in it (perhaps because making the relevant inference would have caused her to lose justification for the premise rather than gain it for the conclusion). Thus if (5) and (6) respectively have the same truth conditions as (10) and (9), then they have different truth conditions from (3\*) and (12). One might therefore conclude that (3\*) does not capture the truth conditions of (5).

However, there is reason to doubt that (5) and (6) respectively do have the same truth conditions as (10) and (9). Consider another sentence of the same form:

$$(13) \quad \text{If an animal escaped from the zoo, it would be a monkey.}$$

The formalization of (13) corresponding to (9) and (10) is (14):

$$(14) \quad \forall x ((\text{Animal}(x) \ \& \ \text{Escapedzoo}(x)) \square \rightarrow \text{Monkey}(x))$$

Consider an elephant; (14) implies that if it had escaped from the zoo, it would have been a monkey. Thus (14) is trivially false. But

(13) is not trivially false; it may well be true. Thus (13) does not have the same truth conditions as (14). For similar reasons, (5) and (6) respectively seem to differ in truth conditions from (10) and (9). Indeed, the very examples used to establish that (3\*) and (12) respectively differ in truth conditions from (10) and (9) tell in favor of (3\*) and (12) rather than (10) and (9) as formalizations of (5) and (6), on at least one reading. Suppose that in the actual world (and, if you like, in all close ones) some farmer owns some donkey and every farmer who owns a donkey beats it; Farmer Giles could have owned this donkey, although he does not own it in the actual world (or in any close one); if he owned it, he would not beat it. In these circumstances, (6) seems to be true on at least one reading, and thereby to have the same truth-value as (12) rather than (9). Similarly, suppose that in the actual world (and, if you like, in all close worlds) someone is Gettier-related to some proposition and everyone who is Gettier-related to a proposition has justified true belief in it without knowledge; that woman could have been Gettier-related to that proposition, although she is not Gettier-related to it in the actual world (or in any close world); if she had been Gettier-related to it, she would have lacked justified belief in it. In these circumstances, (5) seems to be true on at least one reading, and thereby to have the same truth-value as (3\*) rather than (10).<sup>16</sup>

We can formalize (13) along the lines of (3\*) and (12):

$$(15) \quad \exists x (\text{Animal}(x) \ \& \ \text{Escapedzoo}(x)) \rightarrow \\ \forall x ((\text{Animal}(x) \ \& \ \text{Escapedzoo}(x)) \rightarrow \text{Monkey}(x))$$

This deals with the elephant problem. For (15) is true if, had some animal escaped, only monkeys would have escaped; it does not entail that if the elephant had escaped, it would have been a monkey.

The example of (13) also supports the use of universal quantification in the consequents of (3\*) and (12). For suppose that, if some animal had escaped, both a monkey and an elephant would have escaped: then (13) is not true. It is not the case *both* that if an animal escaped it would be a monkey *and* that if an animal escaped it would

<sup>16</sup> As can easily be checked, placing  $\text{Farmer}(x) \ \& \ \text{Donkey}(x)$  in (9) and  $\text{Animal}(x)$  in (14) outside the scope of  $\rightarrow$  makes no serious difference to the argument.

be an elephant. Thus (13) is not equivalent to the result of replacing universal quantification in the consequent of (15) by existential quantification:

$$(16) \quad \exists x \text{ (Animal}(x) \& \text{Escapedzoo}(x)) \square \rightarrow \\ \exists x \text{ (Animal}(x) \& \text{Escapedzoo}(x) \& \text{Monkey}(x))$$

In sloppy terms, what is wrong with (16) as a formalization of (13) is that it does not require the escaping animal with which we started to be a monkey; it is satisfied if some other escaping animal is a monkey. That is not enough to vindicate (13). Analogous points apply to (5) and (6). For purposes of deriving (4), we could have used (17) in place of (3\*):

$$(17) \quad \exists x \exists p \text{ GC}(x, p) \square \rightarrow \exists x \exists p \text{ (GC}(x, p) \& \text{JTB}(x, p) \& \neg K(x, p))$$

But (17) does not entail (5). In sloppy terms, what is wrong with (17) as a formalization of (5) is that it does not require the instance of the Gettier case with which we started to be an instance of justified true belief without knowledge; it is satisfied if some other instance of the Gettier case is an instance of justified true belief without knowledge. That is not enough to vindicate (5). Formalizing (5) as (3\*) avoids this problem.<sup>17</sup>

Henceforth, we assume that (3\*) adequately formalizes the English counterfactual sentence (5).<sup>18</sup> But does (5) adequately express “the Gettier intuition”?

<sup>17</sup> Such truth conditions emerge naturally from accounts that analyze anaphoric pronouns in terms of (not obligatorily singular) definite descriptions (Davies 1981: 166–76, Neale 1990: 180–91); Elbourne (2005) develops a related approach within a framework of situation semantics. It may be less straightforward for alternative approaches to donkey anaphora (such as those based on discourse representation theory or dynamic semantics, for example van Rooij (2006)) to deliver appropriate truth conditions for the relevant sentences: but perhaps it can be done.

<sup>18</sup> For an alternative approach to formalizing the Gettier argument, see Appendix 2.

One might worry that the counterfactual claim (5) overstates the Gettier intuition, just as the claim of strict implication (3) turned out to do. If the actual world happens to contain an abnormal instance of the Gettier case that is not an instance of justified true belief, however many normal instances it also contains that are instances of justified true belief without knowledge, the counterfactual (5) is still false. It is false too if, although the actual world contains no instance of the Gettier case, it happens to be such that if there had been instances, they would have included an abnormal one which was not an instance of justified belief. If it is still possible to have normal instances of the Gettier case which are instances of justified true belief without knowledge, the Gettier intuition might be regarded as still correct, and therefore as not adequately formalized by the false counterfactual (5). Why make the premise of the Gettier argument unnecessarily strong?

We might alleviate the problem by understanding the quantifiers in the formalization (3\*) of (5) as restricted by the conversational context. For example, it might sometimes exclude instances of the Gettier case on Alpha Centauri. However, such restrictions are unlikely to provide a complete solution. For even the contextually relevant domain may happen to betray our expectations.

Here is a simple example. Hank is better at logic than at geography. He wants to refute someone's claim that it is impossible validly to deduce a true conclusion from a false premise. Since he falsely believes that Glasgow is in England, he presents a thought experiment in which "Glasgow is in England or Glasgow is in France" is deduced from "Glasgow is in France." Contextual restrictions do not save Hank. What should we say about this case?

As it stands, Hank's counterexample does not work, and his belief that it works is mistaken. But when the mistake is pointed out, he has no difficulty in repairing it. The easiest repair is simply to substitute "Scotland" for "England." Alternatively, he might stipulate that in his thought experiment Glasgow is in England. One mild disadvantage of the latter stipulation is that it makes the thought experiment depend on an assumption about the contingency of national boundaries which is irrelevant to the logical point at issue. What

would be childish on Hank's part would be to insist that his original thought experiment already constituted a correct counterexample, before he made the stipulation, because he *believed* that Glasgow was in England, and it could have been, so the thought experiment could have been realized in line with his beliefs and, if it had been, it would have been a case of a valid deduction from a false premise to a true conclusion. Although Hank may insist that Glasgow was in England in the case which he had in mind, that was just not the "counterexample" which he actually presented. He spoke falsely when he first said "Someone who infers "Glasgow is in England or Glasgow is in France" from "Glasgow is in France" has validly deduced a true conclusion from a false premise." Similarly, suppose that someone says "Every man in the room is wearing a tie"; I look around, see a man not wearing a tie, misidentify him as Dave (who is in fact wearing a tie), and say "Dave isn't." When it is pointed out to me that Dave is wearing a tie, I deceive myself if I insist that my original reply was correct because the man whom I had in mind was not wearing a tie; that was just not the "counterexample" I actually presented. I spoke falsely when I said "Dave isn't." Even if the audience shares the speaker's false belief that Glasgow is in England or that the man over there is Dave, a third party overhearing the conversation can know that the "counterexample" as it stands is incorrect. For a thought experiment to constitute a counterexample, it is not sufficient that some counterfactual filling out of it, no matter how far-fetched, constitutes a counterexample.

Many philosophers have the common human characteristic of reluctance to admit to having been wrong. We should not distort our account of thought experiments in order to indulge that tendency. Often purported counterexamples fail for accidental reasons and can easily be repaired. To attempt to build into the counterexample in advance all repairs which might conceivably be needed is a futile exercise. It loads the purported counterexample with complexity and in the process weakens it in other respects. The repairs need not articulate qualifications that were in some obscure sense implicit in the thought experiment from the beginning. Rather, they genuinely modify the thought experiment, but the similarity of the new thought experiment to the old one is evidence that the old one was not far wrong.

An example is this. If one is working in the modal system S5, one can weaken the counterfactual premise (3\*) to its mere possibility:

$$(3^{**}) \quad \Diamond(\exists x \exists p \text{ GC}(x, p) \rightarrow \forall x \forall p (\text{GC}(x, p) \rightarrow (\text{JTB}(x, p) \ \& \ \neg K(x, p))))$$

The reason is that in S5, given the necessity of the POSSIBILITY principle, one can reason from  $\Diamond A$  and  $\Diamond(A \rightarrow B)$  to  $\Diamond B$ . For since POSSIBILITY allows the move from  $A \rightarrow B$  to  $\Diamond A \rightarrow \Diamond B$ , it also allows the move from  $\Diamond(A \rightarrow B)$  to  $\Diamond(\Diamond A \rightarrow \Diamond B)$ . But in S5 the application of  $\Diamond$  and  $\Box$  to fully modalized formulas such as  $\Diamond A \rightarrow \Diamond B$  is redundant (modal matters are not themselves contingent), so  $\Diamond(\Diamond A \rightarrow \Diamond B)$  entails  $\Diamond A \rightarrow \Diamond B$ . Consequently,  $\Diamond(A \rightarrow B)$  entails  $\Diamond A \rightarrow \Diamond B$ . In particular, we can deduce (4) from (2) and (3\*\*). Thus one might be tempted to weaken the counterfactual premise to (3\*\*). But that move has its costs too. For it makes thought experiments depend on the soundness of the characteristic principles of S5, whereas the original analysis in terms of (3\*) rather than (3\*\*) involved no such commitment.<sup>19</sup> Moreover, it is strained to attribute the commitment to S5 to people who have never considered the matter when their reasoning can readily be rationalized without it, as before.

Another watering down of the counterfactual premise is to its dual, the negation of the opposite counterfactual:

$$(3^{***}) \quad \neg(\exists x \exists p \text{ GC}(x, p) \rightarrow \neg \forall x \forall p (\text{GC}(x, p) \rightarrow (\text{JTB}(x, p) \ \& \ \neg K(x, p))))$$

Indeed, from (3\*\*\*) one can reason to (4) without invoking (2) as a separate premise.<sup>20</sup> Roughly speaking, (3\*\*\*) says that if the Gettier case had an instance, it *might* be an instance of justified true belief without knowledge, rather than that it *would* be. But (3\*\*\*) falls

<sup>19</sup> Strictly speaking, one must check that that the inference schema from  $\Diamond(A \rightarrow B)$  to  $\Diamond A \rightarrow \Diamond B$  requires the characteristic S5 schema,  $\Diamond \Box A \rightarrow \Box A$ . But substituting  $\neg A$  for A and the contradiction  $\perp$  for B in the inference schema gives the inference from  $\Diamond(\neg A \rightarrow \perp)$  to  $\Diamond \neg A \rightarrow \Diamond \perp$ . Since  $\neg A \rightarrow \perp$  is just the counterfactual equivalent of  $\Box A$  from the previous chapter and  $\Diamond \neg A \rightarrow \Diamond \perp$  is equivalent to  $\Box A$  by normal modal logic, that is tantamount to the inference from  $\Diamond \Box A$  to  $\Box A$ , which is equivalent to the S5 schema, as required.

<sup>20</sup> From the negation of (4) one infers  $\Box \forall x \forall p (\text{JTB}(x, p) \rightarrow K(x, p))$  and thence  $\Box(\exists x \exists p \text{ GC}(x, p) \rightarrow \neg \forall x \forall p (\text{GC}(x, p) \rightarrow (\text{JTB}(x, p) \ \& \ \neg K(x, p))))$  by standard quantified modal logic; the negation of (3\*\*\*) follows by the NECESSITY principle in the previous chapter. Similarly, the negation of (3\*\*\*) follows from the negation of (2). Thus both (2) and (4) follow from (3\*\*\*).

short of normal standards of adequacy for thought experiments. Suppose that a slow-witted philosopher wants to test the hypothesis that the objective probability of a false belief cannot be greater than 99 percent. He imagines himself having bought one ticket in a fair lottery of a thousand tickets with only one winner, believing before the draw that his ticket will lose. He notes that the objective probability of his belief would be greater than 99 percent. However, he has not yet considered whether the scenario is to be one in which his ticket wins. By normal standards, he has not yet determined a counterexample to the hypothesis, although he will have done so once he specifies that in the scenario his ticket wins.<sup>21</sup> Yet presumably the analogue of (3\*\*\*\*) already holds for his unspecific scenario. It is not true that if the unspecific scenario were realized, his ticket would lose – it *might* win. Normal standards of adequacy for thought experiments require something much more like (3\*) than like (3\*\*\*). A similar objection applies to (3\*\*) too.

At the limit, it may be suggested that the role of the Gettier thought experiments is to supply not premises for the conclusion (4) but instead something more like a causal basis for assenting to (4). However, such an undifferentiated account fails to capture what is rational about our rejection of the target analysis. It does not articulate the evidential role of the Gettier case. In most valid deductions, the premises are collectively stronger than the conclusion: that is, the conclusion does not entail every premise. Therefore, they are unnecessarily strong in a purely logical sense. But that sense is not the one that matters. Epistemically and dialectically, the “unnecessarily strong” premises may be exactly what we need. Although their extra strength sometimes leads to trouble, and revision is required, we should not try to cross all such bridges right now, before we come to them; many of them we shall never need to cross.

In any field, arguments are subject to inessential problems of various kinds. Once such problems are identified, they can be fixed without too much difficulty or damage to the original purpose of the argument. We may well be warranted in continuing to attribute the “essential insight” for the argument to its originator, despite his or her minor slips, as we might for the proof of a mathematical theorem. Where reasoning is most explicit, in logic and mathematics, the

<sup>21</sup> For the sake of a simple example, issues about the open future are ignored.

history of mistakes and corrections is often easily documented. Where reasoning is less explicit, as in philosophy, there is more scope for cover-ups. Nevertheless, we should expect that the same process of fine-tuning occurs for philosophical thought experiments as elsewhere. We should not confuse subsequent fallbacks with the original claims. Unnatural formulations such as (3\*\*) and (3\*\*\*) are far more likely to be the fallbacks than to be the original claims. But even when lacunae are identified in a thought experiment, the most likely response in practice is just to add further stipulations to the specification of the case, as it were simply to replace  $GC(x, p)$  by  $GC^+(x, p)$ , so as to preserve the original structure of argument.<sup>22</sup> We resort to the likes of (3\*\*) and (3\*\*\*) only in exceptional circumstances.

One may even wonder whether the move to the counterfactual conditional (3\*) from the strict conditional (3) represents another such fallback. Perhaps: but the question “If there had been an instance of this case, would it have been an instance of justified true belief without knowledge?” seems quite a natural way of articulating what is at stake with a Gettier counterexample. The corresponding questions for (3\*\*) and (3\*\*\*) seem less natural. Moreover, counterfactual questions arise continually in everyday thought, whereas questions of metaphysical necessity rarely arise outside philosophy, so the burden of proof is on those who claim that our initial questions about a hypothetical case are metaphysically modal rather than simply counterfactual in nature. We may, therefore, treat a counterfactual analysis of the arguments underlying philosophical thought experiments as the default. In particular, we may continue to view the Gettier argument as something like the argument from (2) and (3\*) to (4).

## 6

In the original paper, Gettier presents his cases as indicative suppositions. He uses no “subjunctive” conditionals. Although he describes

<sup>22</sup> Merely adding the stipulation that  $x$  and  $p$  constitute a normal instance of the Gettier case is unlikely to solve the problem, for the relevant notion of normality is an epistemological one that violates the supposed neutrality of the initial description of the case.

his target as an attempt “to state necessary and sufficient conditions for someone’s knowing a given proposition” (1963: 121), not as an attempt to analyze the concept of knowledge, we cannot take it for granted that his concern was the metaphysical possibility of justified true belief without knowledge rather than its possibility in some other senses. He wrote before Kripke made the relevant distinctions salient.

Gettier’s intentions aside, why should we not interpret his examples in terms of some non-metaphysical notion of possibility? For instance, we might read the target analysis as the claim that it is conceptually necessary that knowledge coincides with justified true belief. We should then read premise (2) and the conclusion (4) as saying respectively that the Gettier case and justified true belief without knowledge are conceptually possible. If we read (3) as saying that it is conceptually necessary that every instance of the Gettier case is an instance of justified true belief without knowledge, the argument from (2) and (3) to (4) should be valid.

Unfortunately for this reading, the claim that every instance of the Gettier case is an instance of justified true belief without knowledge is unlikely to be conceptually necessary in any useful sense, even if we bracket the general doubts in previous chapters about conceptual modalities. The reason is very similar to that for which we weakened the strict implication premise (3) to the counterfactual conditional premise (3\*). On any reasonable understanding of the phrase “conceptually possible,” it is conceptually possible that some abnormal instance of the Gettier case is not an instance of justified true belief. However, we cannot simply replace the claim of conceptual necessity by the counterfactual premise (3\*). For the argument from (2) and (3\*) to (4) is invalid if the possibility operator in (2) and (4) is understood as conceptual. The POSSIBILITY principle that a counterfactual conditional transmits possibility from its antecedent to its consequent holds for metaphysical but not for conceptual possibility. For instance, friends of conceptual possibility typically think that it is conceptually possible that Hesperus is not Phosphorus but not conceptually possible that Phosphorus is not Phosphorus. But we saw in the previous chapter that the counterfactual conditional “If Hesperus was not Phosphorus, Phosphorus would not be Phosphorus” follows by the logic of identity and counterfactuals from the true identity statement “Hesperus is Phosphorus.”

If the argument from (2) and (3\*) to (4) is to be reworked in terms of conceptual possibility, we need a conditional for (3\*) that stands to conceptual possibility as the counterfactual conditional stands to metaphysical possibility. It is doubtful that the ordinary indicative conditional will do, for “If Hesperus is not Phosphorus, Phosphorus is not Phosphorus” also seems to follow by the logic of identity from “Hesperus is Phosphorus” and the triviality “If Hesperus is not Phosphorus, Hesperus is not Phosphorus.”

Even if we succeeded in cooking up a suitable conditional for (3\*) in respect of conceptual possibility, the reinterpreted argument would show little of philosophical interest. The conclusion would be that it is conceptually possible to have justified true belief without knowledge. That does not refute the hypothesis that knowledge just is justified true belief, of metaphysical necessity, any more than the conceptual possibility of something with atomic number 79 that is not gold refutes the hypothesis that gold just is the element with atomic number 79, of metaphysical necessity. The primary concern of epistemology is with the nature of knowledge, not with the nature of the concept of knowledge. If knowledge were in fact identical with justified true belief, that would be what mattered epistemologically, irrespective of the conceptual possibility of their non-identity. Presumably, if the concept of knowledge were the concept of justified true belief, that identity of concepts would entail the identity of natures, but the converse fails: the non-identity of concepts does not entail the non-identity of natures.

The result of a Gettier thought experiment, interpreted in terms of mere conceptual possibility, would be of significance primarily to theorists of concepts, not to epistemologists. Similarly, the result of a thought experiment in moral philosophy, interpreted in terms of mere conceptual possibility, would be of significance primarily to theorists of concepts, not to moral philosophers. The same would apply to thought experiments in other branches of philosophy. But the use of thought experiments is not confined to the theory of concepts; it flourishes in most branches of philosophy. Consequently, we need an interpretation of them where the possibility at issue is not merely conceptual. The sort of possibility most relevant to the nature of the phenomena under investigation is metaphysical. That fits the approach of this chapter. Nor should we forget how badly the idea of conceptual modality fared under examination in earlier chapters.

The present reflections reinforce the earlier conclusion that it is not a fit instrument for understanding philosophical inquiry.

Related criticisms would apply to the interpretation of philosophical thought experiments in terms of epistemic modalities other than conceptual possibility and necessity. The upshot of a thought experiment in the philosophy of X would be the epistemic possibility (in some sense) of some state of affairs concerning X, not the metaphysical possibility of that state of affairs. That would teach us about the epistemology of beliefs about X, not directly about the nature of X itself. Of course, the epistemology of beliefs about X may indirectly teach us something about the nature of X itself. Indeed, for X = knowledge, the epistemology of beliefs about knowledge is a special case of the philosophy of knowledge, although hardly a representative one. But philosophy does not in general take the diversion of studying X through studying the epistemology of beliefs about X. A more direct approach is feasible. Thus the interpretation of philosophical thought experiments in terms of epistemic possibility is typically inappropriate. Although we may occasionally wish to use them to learn about the epistemology of the object of our study, often we wish to learn more directly about the object of our study itself, in which case a different interpretation of thought experiments is required. The possibility we need then is metaphysical, not epistemic. Thus the non-epistemic approach of this chapter is more widely applicable. Paradigm thought experiments in philosophy are simply valid arguments about counterfactual possibilities.

# Evidence in Philosophy

---

In most intellectual disciplines, assertions are supposed to be backed by evidence. Mathematicians have proofs, biochemists have experiments, historians have documents. You cannot just say whatever you happen to believe. Is philosophy an exception? That hardly fits the emphasis many philosophers place on *arguing* for one's claims. When they cannot provide a deductive argument, they still offer supporting considerations. Often they cite phenomena which, they suggest, their theory best explains: they provide abductive arguments. Indeed, in the last three sentences I gave evidence that philosophers give evidence; so philosophers *do* sometimes give evidence. Of course, philosophers who give evidence that evidence is relevant in philosophy can be accused of begging the question. But let us proceed on the working hypothesis that evidence plays a role in philosophy not radically different from its role in all other intellectual disciplines. Without such a role, what would entitle philosophy to be regarded as a *discipline* at all?

To describe mathematics, biochemistry, and history as evidence-based disciplines is obviously not to subscribe to any extreme foundationalism. Particular appeals to proofs, experiments, and documents can all be questioned. The same goes for philosophy.

In any evidence-based discipline, it is good for an assertion to be consistent with the evidence. The alternative is inconsistency with the evidence, which is bad. Since consistency and inconsistency are relations among truth-evaluatable items, evidence will be treated as consisting of such items, in particular, of propositions. In this sense, the historical evidence is not the physical document itself but various

propositions about it, for example that it is signed “John.” The biochemical evidence is not the experiment as an event but, for example, the proposition that it was carried out with such-and-such results. The mathematical evidence is not the proof as a sequence of steps but, for example, the proposition that the sequence is a correct proof of this claim. This propositional conception of evidence fits the discursive nature of philosophy. When philosophers produce evidence, they produce something truth-evaluable.<sup>1</sup>

Why is it bad for an assertion to be inconsistent with the evidence? A natural answer is: because then it is false. That answer assumes that evidence consists only of *true* propositions. For if an untrue proposition  $p$  is evidence, the proposition that  $p$  is untrue is true but inconsistent with the evidence. Using “fact” for “true proposition,” we may say that evidence consists only of facts. That helps explain the point of conforming one’s beliefs to the evidence.

Although all evidence is true, not all truths are evidence. Some sort of epistemic accessibility is required. Internalists about evidence require the accessibility to be independent of the environment external to the thinker; externalists about evidence reject that requirement. This difference generates a further difference as to what sorts of facts are capable of being evidence. These issues will be considered later.

Since all evidence is true, whatever the evidence entails is also true. The evidence can still support a false proposition non-deductively. If you have not yet heard the result of the lottery, your evidence strongly supports the proposition that your ticket lost, even if in fact it won. Your evidence consists of truths about the lottery available to you at the time.

How can all evidence be true when sometimes the evidence offered turns out to be false? The document was mistranscribed; it was signed “Joan,” not “John.” But the claim that it was signed “Joan” was not really inconsistent with the evidence before the mistranscription was recognized. It was only inconsistent with what was then taken to be the evidence. It was consistent with the fact that the document was transcribed as signed “John.” No evidence was lost when the mistranscription was recognized, and the claim that the document was

<sup>1</sup> Williamson (2000a: 194–200) argues in more detail that propositionality is essential to the functional role of evidence (for the purposes of this chapter, little turns on the choice between sentences and propositions).

signed “Joan” is consistent with the present evidence, so it was consistent with the past evidence. Similarly, biochemists who rely on the misreported results of an experiment are mistaken in saying that part of their evidence for a theory is that the experiment was performed with such-and-such results. Mathematicians who overlook a fallacy in a proof are mistaken in saying that their evidence for the purported theorem is that this sequence of steps is a correct proof of it. Practitioners of any discipline sometimes mistake the extent of their evidence. What is offered as evidence is not always evidence.

Since we can mistake the extent of our evidence, it can be controversial whether a given proposition is evidence. When evidence is not recognized as such, it cannot play its proper role in inquiry. If its status as evidence is controversial, it is not part of the common ground in debate. Relying on a premise one’s opponents have already refused to accept tends to be dialectically useless. They will probably deny that it constitutes evidence; one’s argument will make no headway. As far as possible, we want evidence to play the role of a neutral arbiter between rival theories. Although the complete elimination of accidental mistakes and confusions is virtually impossible, we might hope that whether a proposition constitutes evidence is *in principle* uncontroversially decidable, in the sense that a community of inquirers can always in principle achieve common knowledge as to whether any given proposition constitutes evidence for the inquiry. Call that idea *Evidence Neutrality*. Thus in a debate over a hypothesis  $h$ , proponents and opponents of  $h$  should be able to agree whether some claim  $p$  constitutes evidence without first having to settle their differences over  $h$  itself. Moreover, that agreement should not be erroneous: here as elsewhere, “decidable” means correctly decidable. Barring accidents, if they agree that  $p$  constitutes evidence, it does; if they agree that  $p$  does not constitute evidence, it does not.

One problem for Evidence Neutrality is that the nature of evidence is itself philosophically controversial, as may already be obvious. For example, suppose that a philosophical theory  $T$  entails that every mathematical theorem is evidence, while another philosophical theory  $T^*$  entails that no mathematical theorem is evidence. When proponents of  $T$  debate with proponents of  $T^*$ , whether a given mathematical theorem is evidence is in principle uncontroversially decidable neither positively (since proponents of  $T^*$  are committed to saying that it is not) nor negatively (since proponents of  $T$  are committed

to saying that it is). This objection has the faint air of a self-reflexive paradox, however; perhaps it is an isolated singularity. We turn to more general problems for Evidence Neutrality.

Arguing from the Gettier proposition that the subject in a Gettier case lacks knowledge, I conclude that knowledge is not equivalent to justified true belief. Now I meet someone who thinks the Gettier proposition a mere cultural prejudice, not itself evidence. In this context, it is not in principle uncontroversially decidable that the Gettier proposition is evidence. Thus the only way to satisfy Evidence Neutrality is by ruling that the Gettier proposition does not constitute evidence. To argue that knowledge is not equivalent to justified true belief, I must go back a step to less contentious premises. What can they be? My opponent allows that I *believe* the Gettier proposition, and may even admit to feeling an inclination to believe it too (I am not merely idiosyncratic), while overriding it on theoretical grounds. Thus Evidence Neutrality tempts one to retreat into identifying evidence with uncontroversial propositions about psychological states, that I believe the Gettier proposition and that both of us are inclined to believe it. How much that helps is questionable. For now I face the challenge of arguing from a psychological premise, that I believe or we are inclined to believe the Gettier proposition, to an epistemological conclusion, the Gettier proposition itself. That gap is not easily bridged.

The example depends on no special feature of the Gettier proposition. Any such premise can be questioned and usually is, by skeptics of one sort or another. The dialectical nature of philosophical inquiry exerts general pressure to psychologize evidence, and so distance it from the non-psychological subject matter of the inquiry.

Attempts have been made to close the gap by psychologizing the subject matter of philosophy. If we are investigating our own concepts, our applications of them must be relevant evidence. But this proposal makes large sacrifices for small gains. As seen in earlier chapters, the subject matter of much philosophy is not conceptual in any distinctive sense. Many epistemologists study knowledge, not just the ordinary concept of knowledge. Metaphysicians who study the nature of identity over time ask how things persist, not how we think or say they persist. In such inquiry, the gap between belief and truth is of the same kind as in most non-philosophical inquiry, and the proposal offers little help. Even when one of our own concepts is our subject matter, our inclination to apply it in a given case by no means

guarantees that the application is correct. Cultural prejudices really do sometimes wear the mask of self-evident truth. More generally, the problem with attempts to defend the philosophies of mind and language on the grounds that beliefs about mind and language have a special epistemic status, because they help to constitute their own subject matter, is not just that to extend the argument to other branches of philosophy is to succumb to the usual idealist fallacies. The argument is weak even for the philosophies of mind and language, since our beliefs about our own mind and language can be false for any number of reasons.<sup>2</sup> The gap between belief and truth never completely disappears.

Evidence Neutrality has no more force in philosophy than in other intellectual disciplines: philosophers are lucky if they achieve as much certainty as the natural sciences, without quixotic aspirations for more. If Evidence Neutrality psychologizes evidence in philosophy, it psychologizes it in the natural sciences too. But it is fanciful to regard evidence in the natural sciences as consisting of psychological facts rather than, for example, facts about the results of experiments and measurements. When scientists state their evidence in their publications, they state mainly non-psychological facts (unless they are psychologists); are they not best placed to know what their evidence is? The psychologization of evidence by Evidence Neutrality should be resisted in the natural sciences; it should be resisted in philosophy too. Moreover, not even psychologizing evidence suffices to meet the demands of Evidence Neutrality. For ascriptions of beliefs or inclinations to belief are contestable too, in ways sketched later in this chapter.

Evidence Neutrality is false. Having good evidence for a belief does not require being able to persuade all comers, however strange their views, that you have such good evidence. No human beliefs pass that test. Even in principle, we cannot always decide which propositions constitute evidence prior to deciding the main philosophical issue; sometimes the latter is properly implicated in the former. Elsewhere, I have argued on more general grounds that we are not always in a position to know whether a proposition constitutes evidence (Williamson 2000a: 93–113, 147–83; 2008a). That argument implies

<sup>2</sup> Hintikka (1999) argues that philosophical appeals to “intuitions” were inspired by the paradigm of Chomsky’s linguistics.

the same conclusion, for when it cannot be known whether  $p$  constitutes evidence, it is not in principle uncontroversially decidable whether  $p$  constitutes evidence. Of course, we can *often* decide whether a proposition constitutes evidence prior to deciding the main issue, otherwise the notion of evidence would be useless. But the two sorts of question cannot be kept in strict isolation from each other.

In this respect, philosophy is no different in principle from inquiry in other areas. Since comprehensive physical theories have implications for the reliability of various forms of observation and measurement, they are not neutral as to which reports of such processes constitute evidence. Which axioms of set theory are legitimately assumed in mathematical proofs is itself a mathematical question. Most of the evidence historians cite can be disputed on the basis of perverse conspiracy theories, which are themselves historical theories, however bad. Although philosophy is unusually tolerant of challenges to evidence, no discipline can afford to exclude them altogether, on pain of fatal gullibility.

How much do failures of Evidence Neutrality threaten the conduct of philosophy? From an internal perspective, they make consensus harder. Each of many conflicting theories may be the one best supported by the evidence by its own lights. The role of evidence as a neutral arbiter is undermined. From an external perspective, both the good fortune of being right and the misfortune of being wrong are magnified. If your theory is true, so are its consequences for which propositions constitute evidence; it will be a reliable methodological guide in your further theorizing. If your theory is false, it may have false consequences for which propositions constitute evidence and be an unreliable guide in your further theorizing (if you are very lucky, its falsity is confined to other areas). Although both internal and external effects are damaging, neither is fatal if the failures of Evidence Neutrality are limited enough. The predicament is not special to philosophy, although it may be worse there than elsewhere. It is not in practice fatal to other disciplines; it is not in principle fatal to philosophy.

Unfortunately, the difficulties consequent on failures of Evidence Neutrality are compounded by unawareness of them in much philosophical writing. That unawareness does more than distort philosophers' descriptions of philosophy. It alters their first-order philosophizing, because the regulation of philosophical debate must

be informed by a conception of its nature. For example, the popular but unclear accusation of “question-begging” is leveled on the basis of assumptions about the scope and purpose of philosophical argument.<sup>3</sup> Philosophers under the influence of Evidence Neutrality tend to reject evidence which is not in principle uncontroversially recognizable as such.

These questions are explored below in more detail. They arise with particular urgency from talk of “intuitions.” When contemporary analytic philosophers run out of arguments, they appeal to intuitions. It can seem, and is sometimes said, that any philosophical dispute, when pushed back far enough, turns into a conflict of intuitions about ultimate premises: “In the end, all we have to go on is our intuitions.” Thus intuitions are presented as our evidence in philosophy.

I have heard a professional philosopher argue that persons are not their brains by saying that he had an intuition that he weighed more than three pounds. Surely there are better ways of weighing oneself than by intuition. But such inapposite appeals to intuition should not be dismissed as mere idiosyncratic misjudgments. They are clues to the role of the term “intuition” in contemporary analytic philosophy. Its use may reflect the tacit influence of Evidence Neutrality.

That philosopher knew that if he had simply said that he weighed more than three pounds, rather than that he had an intuition that he weighed more than three pounds, he would have been accused of naïvely begging the question against those who identify persons with their brains. Their theory of personal identity may commit them to denying that he weighed more than three pounds, but not to denying the psychological claim that he had the intuition that he weighed more than three pounds. Thus he used the term “intuition” in an attempt to formulate a psychological premise, not directly about the subject matter of the dispute, which his opponents would concede. Had he been more artful, he might have said that his body weighed more than three pounds, and that he had the intuition that he weighed the same as his body, since they might have conceded both those premises too, and the latter “intuition” has a less empirical flavor.

<sup>3</sup> See Sinnott-Armstrong (1999) for some of the complexities. Naïve attempts to define “begging the question” typically count all deductively valid arguments as question-begging (if you reject the conclusion, you cannot consistently accept the premises).

The point of such maneuvers is primarily dialectical, to find common ground on which to argue with the opponent at hand. The rest of us can be still more confident that he weighed more than three pounds than that he had an intuition that he weighed more than three pounds – he had more chance of deceiving himself or others on the latter point than on the former. But even the dialectical value of such maneuvers is dubious. For if his opponents concede that he has the intuition, they will challenge him to argue from the occurrence of the intuition to its truth: how is he to do that? The simple premise that he weighs more than three pounds at least has the merit of bearing directly on the subject matter of the dispute. Nor need his opponents even concede that he has an intuition that he weighs more than three pounds. They may argue that he is reporting an intuition with some other content, or something other than an intuition.

“Intuition” plays a major role in contemporary analytic philosophy’s self-understanding. Yet there is no agreed or even popular account of how intuition works, no accepted explanation of the hoped-for correlation between our having an intuition that P and its being true that P. Since analytic philosophy prides itself on its rigor, this blank space in its foundations looks like a methodological scandal. Why should intuitions have any authority over the philosophical domain?

## 2

What are intuitions supposed to be, anyway? Let us start by considering a minimalist answer. For David Lewis, “Our ‘intuitions’ are simply opinions” (1983a: x). For Peter van Inwagen, “Our ‘intuitions’ are simply our beliefs – or perhaps, in some cases, the tendencies that make certain beliefs attractive to us, that ‘move’ us in the direction of accepting certain propositions without taking us all the way to acceptance” (1997: 309; he adds parenthetically “Philosophers call their philosophical beliefs intuitions because ‘intuition’ sounds more authoritative than ‘belief’”). If all beliefs or tendencies to belief count as intuitions, then reliance on intuitions is in no way distinctive of philosophy. No scientific progress can be made without reliance on some beliefs and tendencies to belief: simultaneous universal doubt is a dead-end.

In the metaphilosophical debate, that the subject in a Gettier case lacks knowledge is standardly taken as the content of a paradigmatic philosophical intuition. The account of this example in the previous chapter fits indiscriminate characterizations of intuition like Lewis's and van Inwagen's. Our belief in the Gettier proposition (3\*) depends on our capacity to apply epistemological concepts online to encountered instances, our general capacity to apply concepts we can apply online offline too, in the imagination, and our capacity to use such imaginative exercises to evaluate counterfactual conditionals. Far from the brute simplicity which the term "intuition" may suggest, that basis involves complexities absent from the basis of the corresponding judgment about a perceptually encountered Gettier case. For most philosophical purposes, however, the differences between fictional and real-life instances of the Gettier case turned out to be unimportant; what matter are the relevant applications of epistemological concepts, whether offline or online. Nor were those applications especially intimately connected to grasp of the relevant concepts, as some rationalists suggest (Bealer 1998, 2002). Many people grasp the concepts in question without feeling inclined to assent to the Gettier proposition. What they lack is a skill in applying those concepts which goes beyond mere possession. Those who respond correctly to the Gettier case, presented in imagination or perception, do so on the basis of skill in applying the concepts; possessing them is insufficient. None of this encourages the use of the Gettier "intuition" as an exemplar to pick out a special psychological or epistemological kind to which the term "intuition" could helpfully applied.

Epistemologically, the most significant feature of the example may be that many of us *know* the truth of the Gettier proposition. But those trying to demarcate a distinctive category of intuition usually insist that there are false intuitions as well as true ones; they do not project truth from the Gettier example to other cases (for example, Sosa 2006).

George Bealer conceives (rational) intuitions as intellectual seemings (1998: 207; 2002: 73). Background information can defeat our inclination to take perceptual or intellectual seemings at face value. Although we are tempted to believe that one line is longer than the other in the Müller-Lyer illusion, we resist the temptation when we know better. Similarly, the Naïve Comprehension principle for sets,

by which any predicate has a set as its extension, seems true, although we know it to be false, since it is inconsistent by Russell's paradox. But intellectual seemings typically lack the rich phenomenology of perceptual seemings. In its perceptually appearing that something is so, normally in the same event much else perceptually appears too: that various things have various specific shapes and sizes, colors, sounds, tastes, textures, smells . . . Even very primitive sensations have a specific quality of their own. By contrast, in the moment of its intellectually appearing that something is so, often nothing much else intellectually appears. Although mathematical intuition can have a rich phenomenology, even a quasi-perceptual one, for instance in geometry, the intellectual appearance of the Gettier proposition is not like that. Any accompanying imagery is irrelevant. For myself, I am aware of no intellectual seeming beyond my conscious inclination to believe the Gettier proposition. Similarly, I am aware of no intellectual seeming beyond my conscious inclination to believe Naïve Comprehension, which I resist because I know better. I can feel such an inclination even if it is quite stably overridden, and I am not in the least danger of giving way to temptation (just as one can feel the inclination to kick someone without being in the least danger of giving way). Of course, dwelling introspectively for long on *any* belief or inclination to believe has its characteristic phenomenology, but that is the phenomenology of the dwelling, not of what is dwelt upon. These paradigms provide no evidence of intellectual seemings, if the phrase is supposed to mean anything more than intuitions in Lewis's or van Inwagen's sense.

Can we at least restrict intuitions to non-inferential beliefs or inclinations to believe? The belief that one weighs more than three pounds is inferential. So is the belief that there either was or wasn't a cat on this spot exactly five hundred years ago. Yet philosophers often count such beliefs as intuitive, and rejection of them as counterintuitive. If there is a narrower sense of "intuitive," it is often not the operative one when appeal is made in practice to the intuitiveness of some theories as a virtue and the counterintuitiveness of others as a vice.

Does a belief or inclination to believe with an inappropriate causal origin, such as wishful thinking, count as an intuition? We do not want such beliefs or inclinations to believe to carry weight in philosophy. But that is explicable quite independently of whether we classify

them as intuitions. Wishful thinking is as relevant to the epistemology of intuition as misperception is to the epistemology of perception.

Should we restrict philosophical intuitions to those whose basis is grasp of the relevant thought? That is just a variant on the epistemological conceptions of analyticity that were seen to fail in the final section of Chapter 4. The thin grasp of the thought is no basis for assent. The thick grasp of the thought is a basis for assent, but it involves cognitive capacities that are not exclusively conceptual, because they are not necessary for the thin grasp; on such a criterion, intuitions again lose their distinctiveness.

Although we could decide to restrict the term “intuition” to states with some list of psychological or epistemological features, such a stipulation would not explain the more promiscuous role the term plays in the practice of philosophy. This emerges more clearly in appeals to intuition in disputes over actual cases.

Some revisionary metaphysicians deny that, strictly and literally, there are mountains.<sup>4</sup> They deny a proposition of the sort for which G. E. Moore stood up in his defense of common sense (1925). For example, they may argue that although, if there were such a thing as a mountain, it would be a vague object, it is logically impossible for an object to be vague, so there is no such thing as a mountain. Alternatively, they may appeal to ontological economy, and argue that since all the appearances can be explained in terms of the microscopic objects, postulating macroscopic ones in addition is unnecessary and unjustifiable. And so on. The revisionists may concede that microscopic events occur in the joint presence of which it is usual to believe that a mountain is present, but they count that belief false. They hold that although the ordinary use of the word “mountain” has utility, because it registers genuine discriminations between different cases in which different actions are appropriate, it also embodies a mistaken metaphysical theory as to what the difference between those cases consists in (skeptics who doubt that there are mountains may

<sup>4</sup> Van Inwagen (1995) and Horgan (1995) defend related views. They allow that the sentence “There are mountains” may express a truth in some loose or non-literal way, for example when the quantifier is not taken at face value, but in this book “There are mountains” is to be understood strictly and literally. The text presents a metaphysical view of a familiar general type without attempting to follow any one metaphysician in detail.

also be committed to doubting that there are words or beliefs; for the sake of argument we ignore such complications, just as the skeptics tend to do). The claim that there are no mountains is usually regarded as counterintuitive. Even its proponents may concede that it is counterintuitive, arguing that the cost to intuition is worth paying for the overall gain in simplicity, strength, logical coherence, and consonance with science they attribute to their total metaphysical system, which entails the claim. If their system also entails that there could not have been mountains, it contradicts the modal “intuition” that there could have been mountains. But even without the claim of necessity, the non-modal claim that there are no mountains is already counterintuitive as many philosophers use the term, because it contradicts the common sense judgment that there are mountains, for example in Switzerland. Thus the term “intuition” may even be applied to the inferential belief that there are mountains, when based on the belief that there are mountains in Switzerland and elsewhere. Whether or not they agree that there are no mountains, many contemporary metaphysicians would find it philosophically naïve to dismiss a revisionary metaphysical system by appeal to our elementary geographical knowledge that there are mountains in Switzerland. Thus doubts about “intuition” arise for straightforward empirical judgments, even for perceptual judgments: (pointing in the Alps) “Those are mountains.”

Someone could of course stipulate that the only “intuition” in their sense around here is conditional in form: if matter is arranged mountain-wise, then there is a mountain. They would then need to explain what they mean by “mountain-wise.” If they mean *so that it constitutes a mountain*, the purported intuition is an obvious quasi-logical truth: trivially, if matter is arranged so that it constitutes a mountain, then there is a mountain. Perhaps the content of the intuition is supposed to be more like this: if matter is so arranged that *according to the mountain-story* it constitutes a mountain, then there is a mountain. But what exactly is the mountain story? Hard theoretical work is needed to clarify the content of the purported conditional intuition. Once that is done, if it can be, perhaps common sense will be brought to accept the conditional, although it feels more like the conclusion of a plausible argument than the premise. In any case, it lacks the immediate attraction of whatever makes us describe the denial that there are mountains as counterintuitive.

The application of “intuition” and cognate terms in philosophical practice is scarcely more restricted than Lewis and van Inwagen suggest. In general, the objection “That’s only an intuition” is ill-posed in the same way as the objection “That’s only a judgment.” Some judgments are indeed objectionable, but the mere fact that a proposition is judged is not even a *prima facie* reason for doubting it.

Philosophers might be better off not using the word “intuition” and its cognates. Their main current function is not to answer questions about the nature of the evidence on offer but to fudge them, by appearing to provide answers without really doing so. If so, what is really at issue in disputes over the legitimacy of intuitions in philosophy?

### 3

Perhaps skepticism about intuition consists not in skepticism about a special kind of judgment but in a special kind of skepticism about any judgment. That skepticism does not target the distinctive features of perception, memory, testimony, or inference. Rather, it targets our practices of applying concepts in judgment. Call it *judgment skepticism*. For example, it does not question the existence of an external world to which we are causally related in the ways appropriate to perception – at least, not until the concepts of causation and perception themselves come under scrutiny. Indeed, many judgment skeptics are naturalists, their rhetoric scientific. They present themselves as identifying ways in which our conceptual practices need, or may need, revision in the light of scientific advances those practices failed to anticipate. They doubt that we should go on in the same way.

Few judgment skeptics advocate skepticism about all judgments. Total judgment skepticism would result in total intellectual paralysis. Call “judgment skeptics” those skeptical in the way just sketched about some contextually relevant judgments. For example, in a context that concerns folk psychological ascriptions of belief and desire, Paul Churchland and other eliminativists about such mental states are judgment skeptics. In a context that concerns ordinary geographical judgments, Terry Horgan and other eliminativists about

mountains are judgment skeptics. Such skeptics question our standards for applying ordinary concepts both in experience and in thought: the concept of a mountain, the concept of belief, the concept of knowledge, the concept of possibility, the concept of the counterfactual conditional, and so on. Philosophers tend to call judgments “intuitive” when they are considered as the primary targets of judgment skepticism. Thus the term is applied even to the perceptual demonstrative judgment “Those are mountains” or the inferential judgment “There are mountains,” derived by existential generalization – although, for obvious reasons, the primary targets of judgment skepticism are more usually the premises rather than the conclusion of an inference.

Like other skeptics, judgment skeptics construct scenarios to explain how we might make the judgments in question even if they were false. The debunking explanation aims to make massive error a genuine possibility. Scenarios for judgment skepticism are often distinctive in attempting to verify the scientific image of the world while falsifying the manifest image, common sense, or what passes for it in our culture. Sometimes they allow that the ability to apply the key terms of ordinary language (such as “mountain”) in the ordinary way confers an evolutionary advantage, because it helps us communicate to each other genuine but misarticulated differences. The disposition to apply such terms immediately on the basis of casual observation contributes to practical efficiency. Such unreflective discriminations have survival value in harsh environments, where quick decisions are needed. We are here because our ancestors could make them before discovering the true theory of reality. Although the physical theory embedded in our intuitions has to be approximately correct in its predictions over a limited range of practically important cases, we do not expect it to match or even resemble the true physics in representation of the underlying reality. Why should we expect other parts of folk theory to do much better? The cheapest, fastest, and easiest conceptual route for us to making useful discriminations may run through intellectually dirty shortcuts that presuppose false but convenient metaphysics.

In other cases, skeptics may regard a conceptual practice as of merely local value, or even as doing more harm than good. Thus if standards for applying the term “know” vary radically with cultural background, an evolutionary-biological explanation of my current

standard is less plausible.<sup>5</sup> The skeptic may tell a different, more sociological story about the cultural role of knowledge ascriptions, detaching them from their truth conditions. The story might imply that such ascriptions nevertheless fulfill a positive social function to which their cultural variability adapts them. But we can also envisage more sinister stories, on which they serve as instruments of intellectual repression.

Like other skeptics, judgments skeptics ask for independent evidence that favors the piece of common sense at issue over their skeptical hypothesis. The “scientific” flavor of their alternative scenario disguises the resemblance to more traditional forms of skepticism. However, there is one significant difference.

Traditional skeptics argue that we do not know that we are not in a skeptical scenario. They do not positively argue that we are in such a scenario; their point is that we cannot know what our situation really is. For them, the claim that we are in the common sense scenario is no better in epistemic status, but also no worse, than the claim that we are in the skeptical scenario. By contrast, judgment skeptics often argue that we actually are in their skeptical scenario, for example in which there are no mountains, or no beliefs.<sup>6</sup> If they hold that we can recognize that their argument is sound, they must also hold that we can deduce that we are actually in their skeptical scenario. That involves them in no immediate inconsistency, for their skepticism is intended to be partial; they might compare it to skepticism about superstition. Some present their views as superior to “common sense” judgments in compatibility with the results of the natural sciences. They take for granted that those results have some positive epistemic status. Indeed, they often treat them as scientific knowledge. They feel a crisis of confidence in common sense, not in scientific method. For others, it is metaphysical reasoning rather than natural science that trumps common sense.

Despite this more positive aspect of judgment skepticism, judgment skeptics often fall back on traditional skeptical strategies. For instance, they try to put defenders of a piece of common sense into the position

<sup>5</sup> Kornblith (2002) treats knowledge as a natural kind.

<sup>6</sup> Of course, once we stop believing that there are mountains we can no longer be in the full skeptical scenario in which one falsely believes that there are mountains.

of arguing for it over the judgment skeptical scenario from a starting point neutral between the two alternatives, just as skeptics about the external world do.

Judgment skeptics need not puritanically insist that nobody should ever say things like “There are mountains in Switzerland.” Some of their debunking explanations imply that in everyday contexts those are good, useful things to say: outside the metaphysics seminar, utterances of “There are mountains in Switzerland” have more desirable effects than utterances of “There are no mountains in Switzerland.” Discovering the true theory of metaphysics will not change that. Even revisionary metaphysicians can continue to say such things, just as they can continue to say “The sun will rise at 6 a.m. tomorrow.” But, they hold, those things are not strictly and literally true: the sun will not strictly and literally *rise* at 6 a.m. tomorrow; there are not strictly and literally any *mountains* in Switzerland. If we want to think what is *really* true, we must think with the learned; for many purposes it is enough to say what is *to all appearances* true, and speak with the vulgar. We can live most of our lives on the basis of a fiction; only when we take a more scientific attitude are we forced to recognize the fiction for what it is.

For judgment skeptics, appeals to intuition are nothing more than the last resort of dogmatic conservatism, in its desperate attempt to hold back the forward march of scientific and metaphysical progress. But how can such skeptics prevent their arguments for skepticism from applying as far as the sciences themselves?

Judgment skeptical arguments apply to standard perceptual judgments, on which the natural sciences systematically depend: microscopes, telescopes, and other scientific instruments enhance ordinary perception but do not replace it, for we need ordinary perception to use the instruments. If the contents of those perceptual judgments concern ordinary macroscopic objects, they are vulnerable to judgment skepticism about common sense ontology. If so, the empirical evidence for scientific theories is threatened. To assume that the evidence can be reformulated without relevant loss in ontologically neutral terms, in the absence of any actual such reformulation, would be optimistic to the point of naïvety.

Even if that problem could be solved, a more pressing one would remain. Given judgment skeptical arguments, what is the status of scientists’ evidential judgments? For example, suppose that they judge

that a given complex body of evidence of various kinds supports one theory against another, because the former theory explains the evidence better than the latter does. The concept of a better explanation is an informal one, rooted in ordinary ways of thinking, even if scientists' particular applications of it are informed by their background knowledge. A question typical of judgment skepticism arises: what evidence is there that our rankings of explanations are reliable? If the evidence for the hypothesis that our rankings of explanations are reliable is that it provides the best explanation of something else (such as the survival of our species), the charge of question-begging can hardly be ignored. Thus when scientists apply standard concepts of epistemic appraisal, they are not immune to judgment skeptics' styles of argument. In particular, judgment skeptics who judge that our empirical evidence tells against the reliability of some folk theory are vulnerable to judgment skepticism about the elements of folk epistemology on which they are relying.

Although in practice judgment skeptics are often skeptical about only a few judgments or concepts at a time, the underlying forms of argument are far more general. We may suspect that judgment skepticism is a bomb which, if it detonates properly, will blow up the bombers and those whom they hope to promote together with everyone else. But it does not follow that we can dismiss judgment skepticism as self-defeating. That the revolutionary movement would be incapable of establishing a stable new government of its own does not show that it cannot bring the old government down. At worst, judgment skeptics are troublemakers who put on the table arguments we find powerful and in need of a proper response, irrespective of their dubious motives for putting them there.<sup>7</sup>

The similarity between some arguments for judgment skepticism and traditional arguments for traditional forms of skepticism already gives us grounds for suspicion of the arguments for judgment skepticism. If the skeptic about the external world wears the traditional garb of the philosopher while the judgment skeptic dresses up in a scientist's white coat, that should not blind us to the underlying structural similarity of their arguments. A judgment skeptic argues that our evidence is neutral between the ordinary hypothesis that there are mountains and the skeptical hypothesis that there are no

<sup>7</sup> Compare Feyerabend (1978: 143).

mountains, but instead only complex microphysical events the human brain usefully but untruthfully classifies as mountains, and concludes that we cannot know and are not justified in believing that there are mountains. A skeptic about the external world argues that our evidence is neutral between the ordinary hypothesis that there are mountains and the skeptical hypothesis that there are no mountains, but instead only mental states indiscriminable from the inside from perceptions of mountains, and concludes that we cannot know and are not justified in believing that there are mountains. Most people are confident that an argument like the latter for skepticism about the external world is unsound, much less confident as to where exactly it goes wrong. That position is quite reasonable. Similarly, it is fallacious to assume that if one cannot put one's finger on the mistake in an argument for judgment skepticism, one must accept the conclusion, however implausible.

Still, we do want to identify the mistake. Let us therefore consider the epistemological position in more detail, while remembering that the diagnosis of the error in a skeptic's argument may be far less obvious than the fact that it contains an error somewhere.

## 4

Different kinds of skepticism distinguish themselves from each other by questioning some things while leaving others unquestioned. The skeptic about induction grants that all emeralds observed so far were green, in order to question the distinctively inductive step to the conclusion that all emeralds will always be green. The skeptic about deduction grants the premises that if P then Q and that P of an inference by modus ponens, in order to question the distinctively deductive step to its conclusion that Q. The skeptic about testimony grants that someone has said that it was raining, but questions whether she spoke the truth. The skeptic about memory grants that my experience is as of remembering that it was raining, but questions whether I really remember that it was raining. The skeptic about perception grants that my experience is as of seeing that it is raining, so that it visually appears to me that it is raining, but questions whether the experience is veridical. In each case, the skeptic concedes an evidential base, in order to accuse us of going illegitimately beyond it. For the

judgment skeptic, sometimes the only evidential base to hand short of the disputed proposition itself is the conscious inclination to assent to that proposition, to make the judgment.

If judgment skepticism is treated by analogy with skepticism about perception, its evidential base will be described as intellectual seemings, somehow analogous to perceptual seemings. As we saw, Bealer has defended just such an account of intuitions as intellectual seemings. Its intellectually or perceptually seeming to one that  $P$  is a psychological state one can be in whether or not  $P$ , even if the default outcome of being in the state is judging that  $P$ . Whether intellectual seemings are more than conscious inclinations to believe we found reason to doubt.

Skepticism about perception typically narrows one's evidential base to one's present internal mental state. When I can see and hear and feel that it is raining, I suppose my total evidence to include the fact that it is raining, available for assessing hypotheses, for example the hypothesis that the grass will grow. By contrast, the skeptic about perception insists that I have as evidence only the fact that it perceptually appears to me that it is raining, for sometimes what perceptually appears to me is not so. From the fact about my present mental state I am challenged to reason legitimately outwards to the conclusion about my external environment that it really is raining. The skeptic about perception asks by what right I treat the fact that it perceptually appears to me that it is raining as good evidence that it is raining. Judgment skepticism narrows and internalizes our evidential base in a similar way without going as far as skepticism about perception, since typically it treats other people on a par with oneself, and other times on a par with the present. After reading Gettier's article, I suppose my total evidence to include the fact that the subject in a Gettier case lacks knowledge. But the judgment skeptic insists that I have as evidence at most the fact that it non-perceptually appears to me and others that the subject in a Gettier case lacks knowledge, for sometimes what non-perceptually appears to me is not so. From the fact about our mental states we are challenged to reason legitimately outwards to the conclusion that the subject in a Gettier case really does lack knowledge. The judgment skeptic asks by what right we treat the fact that it non-perceptually appears to us that the subject in a Gettier case lacks knowledge as good evidence that the subject in a Gettier case does lack knowledge.

Behind the skeptic about perception's rhetorical question lies an assumption like this: one should be confident that  $P$  (on the basis of perception) only if its (perceptually) appearing that  $P$  is good evidence that  $P$ . Similarly, behind the judgment skeptic's rhetorical question lies an assumption like this: one should be confident that  $P$  (on the basis of common sense) only if its appearing (by the standards of common sense) that  $P$  is good evidence that  $P$ . Call such principles *appearance principles*. They have some initial plausibility. For example, suppose that although whenever I am about to toss a coin either it appears to me that it will come up heads or it appears to me that it will come up tails, such appearances turn out to be correlated no better than chance with the actual results. Its appearing to me that the coin will come up heads is no evidence that it will come up heads. Then I should not take those appearances at face value. Although it appears to me that the coin will come up heads, I should not be confident that it will come up heads.

To discipline the assessment of appearance principles, let us think probabilistically. Say that  $q$  *would be evidence for p* just if  $q$  raises the probability of  $p$ , that is, the conditional probability of  $p$  on  $q$  is higher than the unconditional probability of  $p$ ,  $\text{Prob}(p) < \text{Prob}(p | q)$ .<sup>8</sup> That it appears to me that the coin will come up heads does not raise the probability that the coin will come up heads, so the former proposition would not be evidence for the latter, and I should not be confident that the coin will come up heads on the basis of that appearance, by the relevant appearance principle. More generally, the appearance of  $p$  is *truth-indicative* just if it would be evidence for  $p$  ( $\text{Prob}(p) < \text{Prob}(p | Ap)$ ), and *falsity-indicative* just if it would be evidence against  $p$  ( $\text{Prob}(\neg p) < \text{Prob}(\neg p | Ap)$ , equivalently  $\text{Prob}(p) > \text{Prob}(p | Ap)$ ). An appearance principle implies that one should be confident in  $p$  only if the appearance of  $p$  is truth-indicative

<sup>8</sup> The conditional probability  $\text{Prob}(p | q)$  is usually defined as the ratio of unconditional probabilities  $\text{Prob}(p \ \& \ q)/\text{Prob}(q)$  for  $\text{Prob}(q) > 0$ . The reason for the “would be” is that, in the sense defined, it may happen that  $q$  would be evidence for  $p$  even though  $q$  is itself unknown or even false: the relation between  $p$  and  $q$  is purely conditional. Compare Williamson (2000a: 187). In order to keep the conditional probabilities that are relevant to this chapter uncontroversially well-defined, we allow some metaphysical impossibilities to have non-zero probabilities (for example, according to some judgment skeptics, it may well be metaphysically impossible that there are mountains).

(for specified types of sentence and appearance). A weaker principle says that one should be confident in  $p$  only if the appearance of  $p$  is not falsity-indicative. Note that appearance principles merely purport to give *necessary* conditions for when one should be confident, not sufficient conditions.

On some views, if the prior probability of  $p$  is high enough, we should be confident of  $p$  even if its probability is somewhat lowered by  $Ap$ . The judgment skeptic regards such a defence of disputed philosophical propositions as unacceptably dogmatic, having the advantages of theft over honest toil. Let us concentrate on the unrestricted appearance principles.

What kind of probability should we use to interpret “Prob”? The appearance of  $p$  must not be certain, for if  $\text{Prob}(Ap) = 1$  then automatically  $\text{Prob}(p \mid Ap) = \text{Prob}(p)$ , making truth-indicativeness and falsity-indicativeness uselessly indiscriminate as tests: trivially, the appearance of  $p$  is neither truth-indicative nor falsity-indicative. Thus purely subjective probabilities (credences, degrees of belief) are unsuitable, for the subject may always have been subjectively certain of the appearance of  $p$ . Purely objective probabilities (chances) are also unsuitable, for in a deterministic world with the appearance of  $p$  the appearance of  $p$  is objectively certain. A kind of evidential epistemic probability intermediate between subjective and objective extremes is most relevant.<sup>9</sup> Assume, for the sake of argument, that we have fixed on such probabilities: the discussion below is neutral on their exact nature.

It appears that there are mountains in Switzerland, in a liberal sense of “appears” correlative with the liberal sense of “counterintuitive” in which the claim that there are no mountains in Switzerland is counterintuitive. Presumably, this appearance is truth-indicative, even if significant epistemic probability is assigned to the suggestion of a judgment skeptic that mountains are metaphysically impossible. For there is still a nonzero epistemic probability that mountains are metaphysically possible; conditional on that non-skeptical hypothesis, the appearance that there are mountains in Switzerland surely raises the epistemic probability that there are mountains in Switzerland (for Switzerland might have been a plain), whereas, conditional

<sup>9</sup> Williamson (2000a: 209–37) describes such an intermediate kind of epistemic probability.

on the skeptical hypothesis, the appearance merely leaves the probability unchanged rather than lowering it. Overall, therefore, the appearance that there are mountains in Switzerland raises the epistemic probability that there are mountains in Switzerland.<sup>10</sup> Some philosophically contested “intuitive” propositions are truth-indicative.

However, let SS be the judgment skeptic’s scenario in which it falsely appears that there are mountains in Switzerland, because folk geography misinterprets joint microscopic events as the presence of mountains in Switzerland when in fact mountains are metaphysically impossible. Add to the specification of SS that each trivially necessary condition of there being mountains in Switzerland appears (in the liberal sense) to hold. Since SS is set up as a scenario in which there are no mountains in Switzerland, a trivially necessary condition of there being mountains in Switzerland is that SS does not obtain. Consequently, in SS, it appears that SS does not obtain. Since that is built into the background logic, it is certain, conditional on its not appearing that SS does not obtain, that SS does not obtain. Thus  $\text{Prob}(\neg s \mid \neg A_{\neg s}) = 1$ , where s says that SS obtains. It can be shown to follow that  $\text{Prob}(\neg s \mid A_{\neg s}) \leq \text{Prob}(\neg s)$ , that is, that the appearance that SS does not obtain is not truth-indicative.<sup>11</sup> It is not evidence that SS does not obtain. By the stronger appearance principle, one should not be confident that SS does not obtain.

The judgment skeptic can go further. As already noted, the appearance in question must not be certain, otherwise truth-indicativeness and falsity-indicativeness are trivialized. Thus we may assume  $\text{Prob}(A_{\neg s}) < 1$ . Moreover, those with even the slightest sympathy for judgment skepticism will allow that it is not certain that we are not in SS:  $\text{Prob}(s) > 0$ . These two further assumptions entail that the appearance that SS does not obtain is falsity-indicative: it actually

<sup>10</sup> Formally, where m says that there are mountains in Switzerland and s that mountains are metaphysically impossible, if all the probabilities are well-defined as ratios and  $\text{Prob}(m \mid \neg s) < \text{Prob}(m \mid \neg s \ \& \ Am)$  and  $\text{Prob}(m \mid s) = \text{Prob}(m \mid s \ \& \ Am) = 0$  then  $\text{Prob}(m) < \text{Prob}(m \mid Am)$ . Although there are cases of  $\neg s \ \& \ Am \ \& \ \neg m$ , they are outweighed by cases of  $\neg s \ \& \ Am \ \& \ m$ .

<sup>11</sup> *Proof:* If  $\text{Prob}(\neg s \mid \neg A_{\neg s}) = 1$

then  $\text{Prob}(\neg s \mid A_{\neg s}) = \text{Prob}(A_{\neg s}) \cdot \text{Prob}(\neg s \mid A_{\neg s}) + (1 - \text{Prob}(A_{\neg s})) \cdot \text{Prob}(\neg s \mid A_{\neg s}) \leq \text{Prob}(A_{\neg s}) \cdot \text{Prob}(\neg s \mid A_{\neg s}) + \text{Prob}(\neg A_{\neg s}) \cdot \text{Prob}(\neg s \mid \neg A_{\neg s}) = \text{Prob}(\neg s)$ .

The weaker assumption  $\text{Prob}(\neg s \mid A_{\neg s}) \leq \text{Prob}(\neg s \mid \neg A_{\neg s})$  also suffices.

lowers the probability that SS does not obtain.<sup>12</sup> Its appearing that SS does not obtain is evidence that SS *does* obtain. Therefore, even by the weaker appearance principle, one should not be confident that SS does not obtain.

That there are mountains in Switzerland obviously entails that SS does not obtain. Consequently, one's confidence that there are mountains in Switzerland should be no higher than one's confidence that SS does not obtain: if  $p$  entails  $q$  and subjective probabilities obey the standard probability axioms then the subjective probability of  $p$  is no higher than the subjective probability of  $q$ . By the appearance principle, one's confidence that SS does not obtain should be low. So one's confidence that there are mountains in Switzerland should also be low, even though the appearance that there are mountains in Switzerland is truth-indicative. We therefore face an argument for a sweeping form of judgment skepticism.

The form of argument is not specific to judgment skepticism. It applies equally to skepticism about the external world. We need only replace SS by a skeptical scenario of a more traditional kind. Let  $p$  be a description of the external world acceptable to the judgment skeptic, perhaps in terms of particle physics. Let  $SS^*$  be a scenario in which  $p$  is false but an evil demon makes each trivially necessary condition for the truth of  $p$ , including the truth of  $p$  itself, appear to hold. By the same reasoning as before, it is certain, conditional on its not appearing that  $SS^*$  does not obtain, that  $SS^*$  does not obtain. Thus  $\text{Prob}(\neg s^* \mid \neg A \neg s^*) = 1$ , where  $s^*$  says that  $SS^*$  obtains. It follows that the appearance that  $SS^*$  does not obtain is not truth-indicative; it is not evidence that  $SS^*$  does not obtain. By the relevant appearance principle, one should not be confident that  $SS^*$  does not obtain. The skeptic will further argue that the appearance that  $SS^*$  does not obtain is falsity-indicative; it is evidence that  $SS^*$  *does* obtain. Since  $p$  obviously entails that  $SS^*$  does not obtain, one's confidence in  $p$  should be no higher than one's confidence that  $SS^*$  does not obtain. By the appearance principle, one's confidence that  $SS^*$  does not obtain should be low. So one's confidence in  $p$  should

<sup>12</sup> *Proof:*  $0 < \text{Prob}(s) = \text{Prob}(A \neg s).\text{Prob}(s \mid A \neg s) + \text{Prob}(\neg A \neg s).\text{Prob}(s \mid \neg A \neg s) = \text{Prob}(A \neg s).\text{Prob}(s \mid A \neg s)$  because  $\text{Prob}(s \mid \neg A \neg s) = 0$ . Therefore  $0 < \text{Prob}(s \mid A \neg s)$ , so  $\text{Prob}(\neg s \mid A \neg s) < 1 = \text{Prob}(\neg s \mid \neg A \neg s)$ . Since  $0 < \text{Prob}(\neg A \neg s)$ ,  $\text{Prob}(\neg A \neg s).\text{Prob}(\neg s \mid A \neg s) < \text{Prob}(\neg A \neg s).\text{Prob}(\neg s \mid \neg A \neg s)$ , so the inequality in the previous footnote is strict.

also be low, even if its appearance is truth-indicative. We therefore face an argument for a sweeping form of skepticism about the external world, more specifically, the external world as described in terms the judgment skeptic would accept.

Few judgment skeptics would be consoled by the idea that one's confidence in  $p$  need only be low in contexts in which, since  $SS^*$  has been considered, one must fix a level of confidence in the proposition that  $SS^*$  does not obtain. For they would be unimpressed by a defense of common sense based on the idea that confidence in it is legitimate provided that one refuses to consider their skeptical scenarios. They will insist that head-in-the-sand strategies are futile. They are asking how confident we can be that  $SS^*$  does not obtain, not whether we are capable of ignoring the proposition altogether.

It is unsurprising that if an argument for traditional skepticism works, so does an argument for judgment skepticism. But that is not the kind of success most judgment skeptics seek. They want a more selective skepticism, which for example does not undermine the results of fundamental physics, even though the latter are in the target area for skepticism about the external world. Consequently, they should not use appearance principles as premises in their reasoning, since such principles generate traditional skepticism as well as judgment skepticism. At least in some cases, one can be legitimately confident in a proposition even though its apparent truth is no evidence for its truth, and is even evidence for its falsity.

An observation reinforces that moral. Let  $t$  be any ordinary tautology. The standard probability axioms entail that  $t$  has probability 1, conditional on anything. Then no appearance of  $t$  in any sense is truth-indicative, for  $\text{Prob}(t \mid At) = 1 = \text{Prob}(t)$ . Since  $t$  is also not falsity-indicative, this observation might be met by weakening the requirement of the corresponding appearance principle from truth-indicativeness to lack of falsity-indicativeness. But that misses the intended point of appearance principles. After all, the appearance to me that the coin will come up heads is not falsity-indicative. It does as well as chance, but no better. A different kind of epistemological diagnosis is needed; truth-indicativeness and falsity-indicativeness are just not the relevant criteria.

The problem is not that the definitions of truth-indicativeness and falsity-indicativeness mention only one aspect of appearances, the apparent truth of the proposition  $p$  directly at issue. The arguments

work just the same if we ask whether the totality of appearances (in the relevant sense) would be evidence for  $p$ , given a skeptical scenario  $SS^{**}$  in which  $p$  is false but the totality of appearances matches the actual totality of appearances and all the trivially necessary conditions for the truth of  $p$  appear to hold. For it is certain, conditional on the absence of that totality of appearances, that  $SS^*$  does not obtain. By the same reasoning as before, the totality of appearances is not evidence that  $SS^{**}$  does not obtain, and is even evidence that  $SS^{**}$  does obtain.

Nor is the problem that the arguments were framed in terms of appearances rather than psychological states such as beliefs or dispositions to belief. They work equally well in the latter terms (just substitute  $B$  for  $A$ ).

Rather, the problem concerns a more abstract issue about the structure of confirmation. Let  $e$  be a body of evidence that raises the probability of a hypothesis  $h$  to a value close to 1 without quite making  $h$  certain, so  $\text{Prob}(h) < \text{Prob}(h | e) < 1$ . The material conditional  $e \rightarrow h$  is a logical consequence of  $h$ , and therefore at least as probable as  $h$ ; in fact,  $\text{Prob}(e \rightarrow h | e) = \text{Prob}(h | e)$ . However,  $e$  is evidence against  $e \rightarrow h$ , for  $\text{Prob}(e \rightarrow h) > \text{Prob}(e \rightarrow h | e)$ , simply because  $e \rightarrow h$  is true in all those possibilities which  $e$  eliminates ( $e \rightarrow h$  is a logical consequence of  $\neg e$ ).<sup>13</sup> Clearly, all of this is compatible with a high degree of legitimate confidence in both  $h$  and  $e \rightarrow h$ . Whenever evidence makes some hypothesis more probable than before without making it certain, that evidence makes some logical consequence of that hypothesis less probable than before. Similarly, whenever a hypothesis is certain on some evidence, that evidence makes some logical consequence of that hypothesis no more probable than before (of course, it does not make any such consequence less probable than before, since they all become or remain certain). What this reveals is a fallacy in the tactic of criticizing confidence in a theory by identifying a logical consequence of the theory (not itself a logical

<sup>13</sup> *Proof:*  $\text{Prob}(e \rightarrow h) = 1 - \text{Prob}(e \& \neg h) = 1 - (\text{Prob}(e).\text{Prob}(e \& \neg h | e) + \text{Prob}(\neg e).\text{Prob}(e \& \neg h | \neg e)) = 1 - \text{Prob}(e).\text{Prob}(e \& \neg h | e) > 1 - \text{Prob}(e \& \neg h | e) = \text{Prob}(e \rightarrow h | e)$ . The assumption here that  $\text{Prob}(e).\text{Prob}(e \& \neg h | e) < \text{Prob}(e \& \neg h | e)$  holds because  $\text{Prob}(e) < 1$  (otherwise  $\text{Prob}(h | e) = \text{Prob}(h)$ , contrary to hypothesis) and  $\text{Prob}(e \& \neg h | e) > 0$  (otherwise  $\text{Prob}(h | e) = 1$ , contrary to hypothesis).

truth) whose probability is not raised by the evidence. Call that the *consequence fallacy*.

Consider the deductively valid argument from (1) and (2) to (3):

- (1) Physical events occur that folk geography takes to constitute the presence of mountains in Switzerland.
- (2) If physical events occur that folk geography takes to constitute the presence of mountains in Switzerland, then there are mountains in Switzerland.
- (3) There are mountains in Switzerland.

We may assume that the defender of folk geography is committed to both the premises and the conclusion. In particular, premise (2) is a logical consequence of the common sense conclusion (3) (read the conditional as material). A judgment skeptic may hold that our evidence raises the probability of (1) but not of (2). However, to argue on that basis that, given our evidence, we are not entitled to high degrees of confidence in (2) and (3) is to commit the consequence fallacy.

Similarly, consider the valid argument from (1<sup>\*</sup>) and (2<sup>\*</sup>) to (3<sup>\*</sup>):

- (1<sup>\*</sup>) The Gettier case has features that folk epistemology takes to constitute the subject's lack of knowledge.
- (2<sup>\*</sup>) If the Gettier case has features that folk epistemology takes to constitute the subject's lack of knowledge, then the subject in the Gettier case lacks knowledge.
- (3<sup>\*</sup>) The subject in the Gettier case lacks knowledge.

We may assume that the defender of folk epistemology is committed to both the premises and the conclusion. In particular, premise (2<sup>\*</sup>) is a logical consequence of the common sense conclusion (3<sup>\*</sup>). A judgment skeptic may hold that our evidence raises the probability of (1<sup>\*</sup>) but not of (2<sup>\*</sup>). However, to argue on that basis that, given our evidence, we are not entitled to high degrees of confidence in (2<sup>\*</sup>) and (3<sup>\*</sup>) is again to commit the consequence fallacy.

Finally, consider the valid argument from (1<sup>\*\*</sup>) and (2<sup>\*\*</sup>) to (3<sup>\*\*</sup>):

- (1\*\*) It appears to me that I have hands.
- (2\*\*) If it appears to me that I have hands, then I have hands.
- (3\*\*) I have hands.

As before, the defender of common sense is committed to both the premises and the conclusion. A skeptic about the external world may hold that our evidence raises the probability of (1\*\*) but not of (2\*\*). However, to argue on that basis that, given our evidence, we are not entitled to high degrees of confidence in (2\*\*) and (3\*\*) is once again to commit the consequence fallacy.

The point is doubtless connected to the role of the assumption in some skeptical arguments that knowledge is closed under competent deduction: if I cannot know that I am not a handless brain in a vat that appears to itself to have hands, how can I know that I have hands?<sup>14</sup> However, the arguments in this section have been framed in terms not of knowledge but of legitimate degrees of confidence, conceived as answerable to the standard axioms of probability. In this setting, closure is much less contentious.<sup>15</sup>

## 5

Although judgment skepticism, like other forms of skepticism, easily falls into the consequence fallacy, it would be complacent to assume that it loses all its force once the consequence fallacy has been identified and abjured. We saw in Section 1 the temptation, under the influence of Evidence Neutrality, to conceive the evidence in philosophy as consisting of psychological facts, such as the fact that we believe that there are mountains in Switzerland, not the fact that there are mountains in Switzerland. Since psychological evidence has no obvious bearing on many philosophical issues, judgment skepticism is also encouraged in ways that do not depend on the consequence fallacy. For now the issue is not whether our evidence is evidence for some devious consequence of our theory but whether it is evidence

<sup>14</sup> Seminal works are Dretske (1970), Stine (1976) and Nozick (1981). More recent discussions of closure include DeRose (1995) and Hawthorne (2004); see the latter for more references.

<sup>15</sup> See Williamson 2005c for more discussion of skepticism in relation to truth-indicativeness, and Williamson 2000a: 164-83 for more on traditional skepticism.

for our theory as a whole. And even if our evidence does raise the probability of the whole theory somewhat, is it raised high enough for confidence, in particular, to a higher level than its skeptical alternatives?

Traditional skepticism exploits Evidence Neutrality to achieve a similar psychologization of evidence: only the fact that it appears to me that I have hands is evidence, not the fact that I have hands. How does that happen? Since evidence is true, the false proposition that I have hands is not evidence in a skeptical scenario in which it falsely appears to me that I have hands. Thus the proposition that I have hands is evidence only if I am not in the skeptical scenario. But in the presence of a real or notional skeptic it is contentious that I am not in the skeptical scenario. So it is contentious that the proposition that I have hands is evidence, hence not in principle uncontroversially decidable that it is evidence. Therefore, by Evidence Neutrality, that I have hands is not evidence, even if I am in fact in the common sense scenario in which I have hands and all my perceptual faculties are working properly. Only the proposition that it appears to me that I have hands is evidence. Since both the common sense scenario and the skeptical scenario are consistent with all my evidence, so conceived, the question arises: with what right do I regard the former scenario as more probable than the latter?

Both traditional skepticism and judgment skepticism reflect the tendency of Evidence Neutrality to narrow our evidence base. One result is the uneasy conception many contemporary analytic philosophers have of their own methodology. They think that, in philosophy, ultimately our evidence consists only of intuitions (to use their term for the sake of argument). Under pressure, they take that to mean not that our evidence consists of the mainly non-psychological putative facts which are the contents of those intuitions, but that it consists of the psychological facts to the effect that we have intuitions with those contents, true or false.<sup>16</sup> On such a view, our evidence in philosophy amounts only to psychological facts about ourselves.

<sup>16</sup> A recent example is Brian Weatherson (2003: 27), who, despite showing far more sophistication in these matters than most philosophers do, still assumes that the argument from Gettier cases against the traditional analysis has the premise “Intuition says that Gettier cases are not cases of knowledge” rather than the simpler “Gettier cases are not cases of knowledge.” His considered view may not be the one described in the text.

Nevertheless, they do not want the psychological fact that we have an intuition that  $P$  to be perfectly neutral with respect to the non-psychological question whether  $P$ , for that leads to skepticism about philosophy. If we merely seek the best explanation of our having the intuitions, without any presumption in favor of their truth, we may find a psychological theory to explain them, but how are we to answer the questions about a mainly non-psychological universe that grip many metaphysicians and other philosophers? In explaining why we have intuitions, analytic philosophy has a preference for explanations that make those intuitions true over explanations that make them untrue, but the justification for that preference remains unclear. Even if we have an intuition that the former sort of explanation is better than the latter, why should we give that intuition a special privilege over others by adopting a methodology that assumes its truth? That our evidence in philosophy consists of facts about intuitions and that explanations of those facts on which the intuitions come out true are better (*ceteris paribus*) than explanations on which they do not are themselves epistemological rather than psychological claims. Taken far enough, the psychologization of philosophical method becomes self-defeating. Psychologism is no more a psychological theory than the Pythagorean doctrine that everything consists of numbers is a mathematical theory.<sup>17</sup>

Not even psychological facts really meet the demands of Evidence Neutrality. Whatever Descartes thought, facts about one's own present consciousness are not always cognitively accessible to one. For example, on any reasonable view, intuitions vary in strength. An adequately fine-grained theory of intuitions would have to distinguish weaker ones from stronger ones in evidential impact. If the strength of intuitions is taken into account, the evidence will be recorded in something like the form "I have an intuition of strength  $s$  that  $P$ ." The strength parameter  $s$  will have to be specified according to some common scale, in order to permit the comparisons between the strengths of sometimes conflicting intuitions which the theory of evidence will need to make. But that will give plenty of scope both for

<sup>17</sup> Pust (2001) argues carefully that the following principle is self-defeating: "Aside from propositions describing the occurrence of her judgements,  $S$  is justified in believing only those propositions which are part of the best explanation of  $S$ 's making the judgements that she makes." Contrast Goldman and Pust (1998).

misjudging the strength of one's intuitions and for being accused by others of having done so. After all, philosophers have a powerful vested interest in persuading themselves and others that the intuitions which directly or indirectly favor their position are stronger than they really are. The stronger those intuitions, the more those who appeal to them gain, psychologically and professionally. Given what is known of human psychology, it would be astonishing if such vested interests did not manifest themselves in some degree of wishful thinking, some tendency to overestimate the strength of convenient intuitions and underestimate the strength of inconvenient ones. In trying to compensate for such bias, one may undercompensate or overcompensate; the standpoint of consciousness gives one no privileged access to whether one has succeeded, for bias does not work by purely conscious processes. Its effects are much easier to observe in others than in oneself. A further obstacle to classifying one's intuitions is that some philosophers with a tin ear for natural language seem to misarticulate their own strong intuitions, using forms of words that do not express what they really want. There is sometimes controversy as to whether this has happened. It would be naïve to suppose that all these obstacles can be overcome just by "trying harder." Restricting evidence to psychological facts, even to those about present conscious intuitions, does not satisfy Evidence Neutrality. It is often not in principle uncontroversially decidable whether someone has an intuition of a given degree of strength that P.

Radical eliminativists about the mind are another source of contentiousness. They say "Research in neurophysiology has shown that folk psychology is a false theory; its ascriptions of mental states and acts are never strictly and literally true, however convenient they may have been" (even if they do not believe what they say). At least some of them will classify "S has the intuition that P" and "S has the belief that P" together as ascriptions of folk psychological mental states (perhaps not the same one). On their view (itself a form of judgment skepticism), humans never have the intuition that P. In particular, consistent radical eliminativists will not even concede that their theory is counterintuitive, or that we have the intuition that we have beliefs and desires. To find common ground with radical eliminativists, one must rigorously depychologize one's evidence. I am better off showing them my brain scans than describing my intuitions. For other philosophers, brain scans no more exist than mountains do.

Does a more pragmatic attitude to evidence finesse these difficulties? On a pragmatic view, what permits a fact to serve as evidence in a given context is that it happens to be uncontroversial in that context, not that it is uncontroversial in all contexts, or foundational in any deeper sense. The dialectical standard does not favor the use of psychological facts as evidence in contexts in which such facts are controversial. Currently undisputed non-psychological truths can be used as evidence too. We get by with agreement on particular pieces of evidence without any context-independent standard for evidence. This dialectical conception of evidence makes sense even for a single thinker: in isolation one can still play rival theories against each other in one's head; virtual opponents suffice for much philosophical thinking.

We should not assume too readily that a dialectical standard of evidence is always appropriate. It works well when both sides show moderation and restraint. But the adversarial system of inquiry has limits. By accepting the dialectical standard unconditionally, we lay ourselves open to exploitation by ruthless opponents – such as skeptics. It allows them to rule our best evidence out of court simply by issuing a peremptory challenge to that evidence. A debate conducted in that spirit is unlikely to converge on the truth. The common ground is too narrow to form an adequate evidence base. Testing one's beliefs that way is a dangerous game; we should expect unreliable results. For example, if one uses only premises and forms of inference that a skeptic about perception will allow one, and therefore only premises that are true and forms of inference that are valid even if one is a brain in a vat, one has little prospect of reaching the conclusion that one has hands. But that does not show that we should not be confident that we have hands. To be warranted, confidence need not be recoverable from an impoverished skeptical starting-point. After all, if one uses only premises and forms of inference that skeptics about reason will allow one, one cannot reach the conclusion that there are good reasons. For since such skeptics doubt that there are good reasons, they allow one neither the premise that there are good reasons nor any form of reasoning with which to reach that conclusion from some other starting-point. It would be frivolous to conclude, from that trivial point, that we do not know that there are good reasons. Indeed, even skeptics about reason must deny that conclusion to follow, since they deny that anything follows from anything.

Sometimes, in self-defense, one must abandon skeptics to their fate. Some skepticism, like skepticism about reason, is so radical that it leaves too little unchallenged for what remains as shared evidence to be an appropriate basis for evaluating the claims under challenge. When one is warranted in refusing to play the skeptic's dialectical game, the dialectical standard of evidence becomes irrelevant. In refusing, one does not abandon one's claims to knowledge and reason, for the appropriate standard of evidence is non-dialectical. By that standard, the skeptic's peremptory challenge fails to disqualify the challenged fact as evidence. To neglect such evidence would be to violate the requirement of total evidence.<sup>18</sup> One continues to assert propositions of the disputed kind on the basis of evidence, without expecting to find arguments for them that use only premises and forms of inference acceptable to the skeptic. Since escape from the radical skeptical predicament is impossible, one must take good care not to get into it in the first place.

Is this attitude a legitimate response to judgment skepticism? For instance, may one take the fact that the subject in a Gettier case lacks knowledge or the fact that there are mountains in Switzerland as evidence, even though the judgment skeptic challenges one's right to such evidence? In reaching one's views, one does not restrict oneself to premises and forms of inference acceptable to judgment skeptics, for one regards their restricted evidence base as too willfully impoverished to constitute a reasonable starting-point for inquiry. Such skeptics have not shown that the facts they allow as evidence are really more certain than the facts they disallow. In particular, it is quite insufficient for them to point out that it is possible to judge that there are mountains in Switzerland even if there are no mountains in Switzerland, for a parallel objection can be made to any evidence worth having in the sciences.

Even if (let us pretend) facts about our intuitions were in some sense more certain for us than all other facts, it would not follow that we should restrict our evidence to facts about our intuitions. For the extra information in a wider evidence base may be worth a cost in reliability. If logical truths were more certain than all other facts,

<sup>18</sup> “[I]n the application of inductive logic to a given knowledge situation, the total evidence available must be taken as a basis for determining the degree of confirmation” (Carnap 1950: 211; compare Hempel 1965: 63–7). See also Williamson (2000a: 189–90).

it would not follow that we should restrict our evidence to logical truths: that would eliminate most of our knowledge. It would be skepticism about everything except reason. Similarly, if facts of some other special kind were more certain than all other facts, it would not follow that we should restrict our evidence to facts of that special kind.

Isn't this short way with the judgment skeptic contrary to the open spirit of philosophical discussion? The skeptic has thoughtful, recognizably philosophical concerns: don't they deserve a fair hearing? How can they be given such a hearing if the very propositions the skeptic challenges are taken as evidence? Skeptics of any principled kind can indeed expect more tolerance in philosophy than in other disciplines. One can discuss their skepticism with them without stepping outside the bounds of philosophy. In talking to them, it is futile to offer for their acceptance arguments with premises they have already refused to accept. In particular, it seems unphilosophical to refuse to discuss judgment skepticism with its proponents. In conversation with them, it is dialectically pointless, rude, to offer as evidence propositions one knows they do not accept. But the issue remains: what implications, if any, does the outcome of such a conversation have for the epistemic status of belief in the propositions the skeptic questions? Faced with a skeptic about reason, or everything except reason, many philosophers would be willing to start a conversation, out of politeness, curiosity, competitiveness, or the desire to save a soul. But their inability to achieve a dialectical triumph over such a resourceful opponent does not oblige them to become skeptics about reason, or everything except reason, themselves. There is no bad faith in continuing to claim (and have) knowledge of the contested truths. For the anti-skeptic is not obliged to treat dialectic as the measure of all things. Indeed, the claim that dialectic is the measure of all things faces self-defeat, for it cannot triumph dialectically over its denial; even if it appeared to be getting the better of the argument, would not taking that to establish its truth beg the question? Similarly, even if one cannot establish dialectically, in dispute with a judgment skeptic, that the subject in a Gettier case lacks knowledge or that there are mountains in Switzerland, without bad faith one can still claim to know that the subject in a Gettier case lacks knowledge or that there are mountains in Switzerland, and use those facts as evidence.

What prevents astrologers from using this approach to defend astrology by arguing that the fact that astrological predictions have an excellent track record constitutes good evidence for astrological theory? Nothing prevents astrologers from *saying* such things, although they will presumably be speaking falsely, since astrological predictions have no such excellent track record. Similarly, nothing prevents astrologers from *saying* that astrology meets the strictest methodological standards of natural science, although again they will be speaking falsely. In both cases, there will be excellent evidence that they are speaking falsely, which they will not accept as evidence of that. There is a persistent temptation to assume that a good account of methodology should *silence* astrologers and other cranks, by leaving them in a position where they can find nothing more to say. That assumption is naïve. They always find more to say. Of course an account of methodology should specify respects in which good intellectual practices are better than bad ones. But that does not mean that if devotees of a bad intellectual practice endorse the account, they will abandon the practice; more likely they will convince themselves that their practice triumphantly conforms to its precepts. No methodology is proof against misapplication by those with sufficiently poor judgment.

None of the foregoing arguments provides any guarantee that judgment skepticism is not correct for some types of judgment; “common sense” is sometimes wrong. But if it is accepted in such cases, that should be on the basis of evidence specific to those types of judgment, not on the basis of general skeptical fallacies.

## 6

Our evidence in philosophy consists of facts, most of them non-psychological, to which we have appropriate epistemic access. Consequently, there is a one-sided incompleteness to descriptions of philosophical methodology, and attempts to justify or criticize it on that basis, if formulated in terms neutral over the extent of that evidence. For instance, in describing some philosophers as believing or having the intuition that *P*, one fails to specify whether their evidence includes the fact that *P*.

A simple attempt to justify common sense as a starting point for philosophy on the basis of such a neutral description appeals to the

principle of *Epistemic Conservativism*: one has a defeasible right to one's beliefs, which may be defeated by positive reasons for doubt, but not by the mere absence of independent justification.<sup>19</sup> Thus one's belief that there are mountains in Switzerland gives one the defeasible right to rest arguments on the premise that there are mountains in Switzerland. Whether or not the belief constitutes knowledge, it confers the right.

Our beliefs are what we start from, the boat we find ourselves in. Even if we can progressively replace them, we cannot distance ourselves from all of them at once, for we have nowhere else to stand. Epistemic Conservativism elevates the practical necessity of starting from where one is, wherever that is, to normative status, subject to the proviso on defeaters. Although the principle is not perfectly neutral on the epistemic status of the belief, since the notion of a defeater is epistemologically normative, it is neutral on how much evidence, if any, the subject has. Justifying a philosophical method by appeal only to Epistemic Conservativism ignores crucial epistemological distinctions concerning the relevant beliefs: it is like justifying scientific methodology without giving any information as to what evidence is required in its application. Even if Epistemic Conservativism is true, it is radically incomplete as a basis for an account of the epistemic status of philosophical beliefs.

If philosophical "intuitions" are simply beliefs, they fall within the domain of Epistemic Conservativism. That is less clear if "intuitions" include inclinations to belief. Someone inclined to believe  $p$  may nevertheless not believe  $p$ ; inclinations conflict. This difference matters for Epistemic Conservativism.

Justin has been brought up to believe that knowledge is equivalent to justified true belief. He is confronted for the first time with a Gettier case. He might have immediately and confidently judged that the subject has justified true belief without knowledge, and abandoned his old belief that knowledge is equivalent to justified true belief. Presumably, Epistemic Conservativism would then have switched sides and started supporting the new belief that knowledge is not equivalent to justified true belief. Instead, Justin is more cau-

<sup>19</sup> See Harman (1986: 29–42) for a defense of epistemic conservatism, and Vahid (2004) for a recent critical survey of its varieties. For simplicity and generality, subtleties in the formulation of the principle have been glossed over.

tious, not wanting to assent too readily to anything tricky. Although he is consciously inclined to judge that the subject has justified true belief without knowledge, he does not immediately give in to that inclination or abandon his ingrained belief that knowledge is equivalent to justified true belief. Does Epistemic Conservativism counsel abandoning his ingrained belief in this situation? If Justin is asked “What reason have you to doubt your analysis?,” he cannot answer “The subject in this possible case has justified true belief without knowledge,” since he does not yet believe that. He must say something else. The answer “I am inclined to believe that the subject in this possible case has justified true belief without knowledge” would be relevant if the function of the prefix “I am inclined to believe that” were to signal tentative assent to what follows, but Justin’s commitment to his analysis inclines him to resist even tentative assent to a putative counterexample. If the function of the prefix “I am inclined to believe that” is instead to report his psychological state of being inclined to believe the proposition expressed by the embedded sentence, as its literal compositional semantics suggests, the relevance of that answer to the original question is far from obvious, for he has not yet assented even tentatively to a counterexample.

Can Epistemic Conservativism be extended to the claim that one has a defeasible right to believe whatever one is inclined to believe? Such an extension is less clearly motivated than the original principle by the idea that, since one must start from where one is, one has at least a defeasible right to be there. A right to be where I am is a right to have the beliefs and inclinations I have. That does not obviously include a right to follow those inclinations to new places, especially when the beliefs I already have imply that those are bad destinations, for example, when the inclinations are to believe things inconsistent with what I currently believe. As Gettier counterexamples show, intuition can be revolutionary as well as conservative. If I currently believe  $p$ , I am currently committed to the belief that any inclination to believe something inconsistent with  $p$  is an inclination to believe something false. I am not committed to the beliefs I am merely inclined to have in the way I am committed to my current beliefs. I am merely inclined to commit myself to them in that way. After all, a right to be where I am is of limited practical use unless it involves a right to stay where I am, to continue believing, at least for a while, what I currently believe.

Many philosophers recognize their philosophical activity in the more dynamic notion of reflective equilibrium, described by Nelson Goodman and John Rawls.<sup>20</sup> Our initial set of general theories and particular intuitions is inconsistent; each side is revised in the light of the other, by an iterative process, until they are brought into harmony. There is a debate whether the beliefs that emerge from this process are thereby justified. But a prior question is whether such descriptions of the process yield an adequate conception of a philosophical method, good or bad. The question is not whether philosophers engage in the mutual adjustment of general theory and judgments about specific cases – they manifestly do – but whether such descriptions of it are sufficiently informative for epistemological purposes.

A process generally acknowledged as at least superficially analogous to the attainment of reflective equilibrium in philosophy is the mutual adjustment of theory and observation in natural science.<sup>21</sup> Imagine a description of it in which the word “observation” is used simply as a label for judgments with non-general content, irrespective of origin; it ignores the perceptual process. Such a description misses the point of the natural scientific enterprise. It provides no basis for an epistemological assessment. The nature of scientists’ evidence has been left unspecified. Similarly, one has no basis for an epistemological assessment of the method of reflective equilibrium in philosophy without more information about the epistemological status of the “intuitions.” In particular, it matters what kind of evidence “intuitions” provide. The previous account of thought experiments is consistent with the idea that the Gettier proposition and its like are evidence. Indeed, since real life counterexamples will sometimes do in place of imaginary ones, observed facts are sometimes relevant evidence. Talk of reflective equilibrium fails to address such issues.

<sup>20</sup> See Goodman (1955: 65–8) and Rawls (1951, 1971: 20). David Lewis (1983a: x) describes philosophers’ task as the identification of such equilibria. Two recent critiques of the method are Cummins (1998) and Stich (1998); a recent defense is DePaul (1998).

<sup>21</sup> For such an analogy see Rawls (1951).

One factor obscuring the descriptive inadequacy of standard accounts of reflective equilibrium is the already noted tendency to conceive evidence in philosophy as the mere having of “intuitions”: it is easy to slip into the illusion that our epistemic access to such psychological facts is unproblematic. Thus attention is distracted away from the epistemic status of the “intuitions” themselves. Even if we revise an “intuition,” our evidence may still include the fact that we had it. But the epistemic status of the original “intuition,” however much the model ignores it, must be relevant to the epistemic value of revising general theories in line with its content.

The reflective equilibrium account, as usually understood, already assigns a proto-evidential role to at least one kind of non-psychological fact. For it treats philosophers as relying on logical relations between theories and intuitions, in particular their consistency and inconsistency. Can one retell the story in purely psychological terms, with beliefs about logical relations in place of actual logical relations? That move is doubly problematic. It reduces explanatory power unless the assumption is added that beliefs about logical relations are reliable, for otherwise the account no longer explains any tendency to bring theory and intuition into mutual consistency, but at best a tendency to believe that one has done so. Moreover, the beliefs about logical relations are explanatorily redundant. Consider the theory (4) and the “intuition” (5):

- (4) Every F is a G.
- (5) This F is no G.

In order to explain, without appeal to the inconsistency of (4) with (5), why philosophers do not simply retain both, we merely say that they believe (6):

- (6) (4) and (5) are jointly inconsistent.

Philosophers do not in fact fix belief in all of (4), (5), and (6). But the envisaged strategy does not understand that in terms of a proto-evidential role for (7):

- (7) (4), (5), and (6) are jointly inconsistent.

It no more assumes that (7) is evidence than it assumes that (6) is. To invoke the fact of belief in (7) as evidence is merely to take another backwards step on an infinite regress. But if the strategy relies on a brute unwillingness to believe all three of (4), (5), and (6), it might as well have relied on a brute unwillingness to believe both of (4) and (5) in the first place; they are already inconsistent. Without proto-evidential backing from the inconsistency of (4) and (5), the unwillingness to believe both of (4) and (5) looks irrational.

If the reflective equilibrium story assigns a proto-evidential role to some logical facts even though all logical facts are philosophically contestable, as we saw in earlier chapters, why not allow a similar role to other philosophically contestable facts? If no other philosophically contestable facts can play such a role that is something we need to know, and have not yet been given any good reason to believe. If other philosophically contestable facts can play a proto-evidential role, that too is something we need to know and which the reflective equilibrium story leaves unacknowledged.

To say that mathematicians or biochemists or historians strive to bring their opinions into equilibrium would be sadly inadequate as even a summary description of their method of research. It omits the constraining evidence that makes their opinions worth listening to, their research worth funding. Is philosophy so different that in its case such a description will suffice? If so, it should give up any claim to be an evidence-based discipline. Such pessimism is unwarranted once we accept the contestability of evidence. Thought experiments do provide evidence, in the shape of mainly non-psychological facts. That philosophers sometimes disagree as to what evidence they provide is only to be expected.

# Knowledge Maximization

---

In philosophy, as elsewhere, one can easily experience conflict between one's role as a believer and one's role as an appraiser of oneself as a believer. I cannot *simply* regard my belief in  $p$  as a psychological phenomenon. For  $p$  implies that  $p$  is true, and therefore that whoever believes  $p$  does so truly. In believing  $p$ , I am committed by that implication to the belief that a belief in  $p$  is true, and that to continue believing truly on the matter I must continue believing  $p$  (if the truth-value of  $p$  is atemporal). Similarly,  $p$  implies that its negation  $\neg p$  is false, and therefore that whoever believes  $\neg p$  has a false belief in  $\neg p$ . In believing  $p$ , I am committed by that implication to the belief that a belief in  $\neg p$  is false.<sup>1</sup> Neutrality is not an option for believers. One is bound to think any given belief of one's own superior in truth-value to the contrary beliefs of others. But sometimes we step back from our beliefs and regard them as psychological phenomena on a par with the beliefs of others, in equal need of both psychological explanation and epistemological criticism. I may see my beliefs as the product of my social and cultural background, your beliefs as the product of your social and cultural background, and wonder what objective reason there is to prefer mine to yours. As argued in the previous chapter, that third-person stance can involve a refusal to take crucial knowledge seriously, just because someone disputes it; sometimes we must take a first-person present tense stance. But

<sup>1</sup> The exact status of the implications depends on delicate issues about disquotational principles for truth and falsity, but whatever their outcome the point in the text will hold in some form.

sometimes the third-person stance is the right one to take. This chapter explores some general aspects of the tension between one's role as a believer and one's role as an appraiser of oneself as a believer in philosophy.

Some anti-skeptical commitment is built into the role of believer. If I believe  $p$ , I am thereby committed to the belief that I do not falsely believe  $p$ . This commitment can be generalized in two ways. First, the content of the hypothetical commitment can be generalized. If I believe  $p$ , I am thereby committed to the belief that no one falsely believes  $p$ . Similarly, given that propositional truth is atemporal, if I believe  $p$ , I am thereby committed to the belief that I shall never falsely believe  $p$  (since propositional truth is not amodal, there is no corresponding modal generalization: if I believe  $p$ , my commitments may allow that  $p$  could have been false and still believed). Second, the whole conditional can be generalized, on personal, temporal, and modal dimensions. Necessarily, anyone who ever believes  $p$  is thereby committed to the belief that they do not falsely believe  $p$ . All these generalizations can be combined: necessarily, anyone who ever believes  $p$  is thereby committed to the belief that no one ever falsely believes  $p$ .

Nevertheless, this anti-skeptical commitment is very limited. For if I believe  $p$ , my commitments may allow that just about everyone else falsely believes  $\neg p$  at all times in all circumstances, that I falsely believe  $\neg p$  at just about all other times in all circumstances, and that I would now have falsely believed  $\neg p$  in just about all counterfactual circumstances: true belief with respect to  $p$  in the current case contrasts with error on the same question in just about all other cases. I might take skeptical scenarios to prevail almost everywhere while insisting that I happen not to be currently in one. Such a response to skepticism would be unimpressive, perhaps unstable. The admitted frequency of skeptical scenarios in nearby situations constitutes an urgent reason for doubting one's own beliefs. One should beware of regarding oneself as too happy an exception to sadly general trends. Sometimes the tension between one's role as a believer and one's role as an appraiser of oneself as a believer becomes unbearable, and the belief in question is abandoned.

Few of us regard ourselves as highly exceptional in having currently escaped the worst scenarios for skepticism about perception. We think them rare in worlds like ours. We find the brain in a vat

scenario far-fetched; while dreams are common, dreams with the sustained coherence of waking life are very rare. The environment as we perceive it is full of creatures in regular perceptual contact with it. No special luck or skill is needed to avoid envatment: it has never been a big danger for humans. Of course, skeptics will say that such claims about our environment merely beg the question; their truth is part of what is at stake. But the claims were not addressed to skeptics, in a futile attempt to persuade them out of skepticism. Instead, they figure in our appraisal of skeptical arguments, from our current non-skeptical point of view.<sup>2</sup> Not yet having suspended our ordinary beliefs, we must decide whether the acknowledged bare metaphysical possibility of skeptical scenarios gives us good reason to suspend those beliefs – not just momentarily in an epistemology seminar, but for the rest of our lives. Most of us find the reason inadequate. Bare possibilities of error, however picturesque, constitute no imminent threat; the threat is not nearly urgent enough to warrant the drastic and costly precautions skeptics recommend. For most purposes, we do not take the skeptical possibilities seriously.

Our tendency to ignore skeptical possibilities is not explained by their making no practical difference; many of them make such a difference. If you are a brain in a vat, not really interacting with other people, much of your altruistic behavior is futile. Again, in some skeptical scenarios you feel unremitting horrible pain for years, starting tomorrow, unless you immediately do what appears to you exactly like going out and buying ten copies of the same newspaper: I bet you do not take even that elementary precaution. Of course, in other skeptical scenarios you feel unremitting horrible pain for years, starting tomorrow, if you immediately do what appears to you exactly like going out and buying ten copies of the same newspaper. If one takes all possibilities equally seriously, they tend to cancel each other out for practical purposes. But that does not imply that we are left back where we were before skeptical possibilities occurred to us. If everything except present consciousness is utterly unknown, why not simply indulge in sweet dreams?

For the thorough skeptic, that you have hands is no more probable (epistemically) than that you are in a skeptical scenario in which you merely appear to have hands: will you therefore reject a bet on which

<sup>2</sup> Compare Nozick (1981: 167).

you win 10 euros if you have hands and lose 100 euros otherwise on the grounds that its expected utility is negative, since  $10/2 - 100/2 = -45$ ? If skepticism makes you doubt the enforceability of the bet, that is no reason to accept it. Surely it is a good bet, even when you happen to be in an epistemology seminar. We ignore radical skeptical possibilities in practice, even when they are drawn to our attention, because we do not rate them as epistemically serious possibilities. We make that epistemic assessment from our non-skeptical perspective.

When we judge that in our world radical skeptical scenarios present no imminent danger to anyone, we do so on the basis of our own beliefs, but that judgment depends on the specific content of those beliefs; it is not automatic. We have a rich conception of ourselves and our environment on which brains in vats are very far-out physical possibilities, and even long-term coherent dreams are highly unlikely. That conception also enables us to give specific answers to the question “How do you know?” as it arises on specific occasions, for example by indicating relevant processes of perception, memory, testimony, and inference, although of course the conception need not figure among premises from which the more specific knowledge was inferred, since the latter need not have been inferred at all. None of this amounts to a detailed dissection of the flaws in particular skeptical arguments. Rather, it provides the appropriate background to our confidence that such flaws must be there.

How imminent a threat do scenarios for judgment skepticism pose? Skepticism about perception starts with actual perceptual errors and imaginatively radicalizes them until it reaches brains in vats. Similarly, judgment skepticism starts with actual errors about witchcraft, oracles, and magic and imaginatively radicalizes them until it reaches the nonexistence of mountains. In both cases, there is a trade-off between how remote the skeptical scenarios are (judged from our current perspective) and how far-reaching a skepticism they motivate. The set of very close possibilities motivates only a very limited skepticism; a wider range of possibilities motivates a more general skepticism. The closer the possibility, the more seriously it deserves to be taken. For skepticism about perception, we know at least roughly what makes the more radical scenarios remote, the enormous practical obstacles to setting up all the requisite causal mechanisms, not to mention the shortage of motivation for doing so. For judgment skepticism, what corresponds to those obstacles? Do we even believe

that the actual world is not full of apt scenarios for judgment skepticism?

Suppose that most ordinary beliefs in most other cultures are false, because somehow laden with false theories.<sup>3</sup> Then the possibility that, for similar reasons, most ordinary beliefs in our own culture are also false is too close to home to be dismissed as fanciful or far-fetched. Judgment skepticism gets a grip. A satisfying response would put such skeptical scenarios far from other cultures, not just far from one's own.

Given empirical evidence for the approximate intertranslatability of all human languages and a universal innate basis of human cognition, we may wonder how "other" any human culture really is. If we believe  $p$  and believe that others believe  $p$  too, then we are committed to the belief that the others' belief in  $p$  is true. But if human beliefs tend to be true merely as an accidental by-product of our DNA, and other galaxies are rife with nonhuman persons most of whose beliefs are false, because laden with false theories, then scenarios for judgment skepticism are still dangerously close to home. Even if such scenarios are rare or absent in the actual universe, but only by good luck, it remains uncomfortable for opponents of judgment skepticism. If we are to refuse in good conscience to take seriously the radical scenarios for judgment skepticism, we must do so from a perspective on which there is a quite general tendency for beliefs to be true. Anything less than that will look like special pleading on our own behalf. But why should there be any such tendency? What we believe is one question, what is true another.

## 2

Some naturalists argue on evolutionary grounds that beliefs tend to be true, for creatures with too many false beliefs are unfit to survive. True beliefs tend to cause one to get what one wants in a way in which false beliefs do not. Truth conduces to success. That is not to deny that some false beliefs have survival value; the suggestion is only that on the whole truth is more conducive than falsity to survival. Since we are arguing from our current perspective, on which our

<sup>3</sup> For present purposes, how finely cultures are individuated matters little.

world is governed by regularities extending over past, present, and future, we need not worry overmuch about scenarios for inductive skepticism on which generalizations with only true instances up to some future time  $t$  have false instances thereafter (in any case, judgment skepticism is not skepticism about induction). We can take past success as some guide to future success.

How do true beliefs tend to cause success in action? This principle seems central to the nature of belief and desire:

- (1) If an agent desires that  $P$ , and believes that if it does  $A$  then  $P$ , then *ceteris paribus* it does  $A$ .

The “*ceteris paribus*” clause in (1) covers possibilities of irrationality, alternative means to the same end, countervailing desires, and so on. If an agent desires that  $P$ , believes that if it does  $A$  then  $P$ , and does  $A$ , then  $P$  if the belief is true, so its desire is realized. If its belief is not true, then it may well not happen that  $P$ . Of course, that  $P$  may not help the agent if it is not good for the agent that  $P$ . The argument might therefore be taken to support a stronger conclusion: that evolution favors creatures who both believe what is true and desire what is good for them. “Good for them” here means good for them collectively, since evolution sometimes favors altruistic behavior which benefits one’s relatives to one’s individual disadvantage; for simplicity, this qualification is left tacit in what follows.<sup>4</sup>

An agent has some idea of the act  $A$  in believing that if it does  $A$  then  $P$ . If it does  $A$  without believing itself to be doing so, then the natural link between antecedent and consequent in (1) is broken. For example, if you go north while believing that you are going south,

<sup>4</sup> If for them to desire that  $P$  were for them to believe that it is good for them that  $P$ , the tendency to desire what is good for them might be subsumed under the tendency to believe what is true. However, in whatever sense of “good for them” evolution can be assumed to favor creatures that get what is good for them, for them to believe that it is good for them that  $P$  seems to be neither necessary nor sufficient for them to desire that  $P$ . For example, they may believe that it is good for them in that sense that there be a cull of the unfit without desiring one, and they may desire that cigarettes be more readily available without believing that it is good for them in that sense that cigarettes be more readily available. But if in some relevant sense desiring that  $P$  can be equated with believing that it is good that  $P$ , so much the better for the argument in the text.

your action is not explained just by your desire to reach the oasis and belief that if you go north then you will reach the oasis, (1) notwithstanding. Perhaps the explanation is that, in addition, you desire even more strongly to avoid your enemy and believe that he is at the oasis. Although such examples do not refute (1), since the “*ceteris paribus*” clause absorbs their shock, they indicate that the rationale for (1) takes for granted that beliefs about what one is doing tend to be true, which is a special case of the very phenomenon that we are trying to understand. Therefore, in order not to assume what needs to be explained, let us revise (1) thus:

- (2) If an agent desires that  $P$ , and believes that if it does  $A$  then  $P$ , then *ceteris paribus* it acts so that it believes that it does  $A$ .

A natural variant of (2) would have “on the intention to do  $A$ ” in place of “so that it believes that it does  $A$ .” The argument below could be reformulated in terms of this variant, but for simplicity let us stick with (2), to minimize the number of types of propositional attitude under consideration.

Given that you want to avoid your enemy, and believe that if you go south then you will avoid him, (2) helps explain why you act so that you believe that you go south, even though in fact you go north. But the reason for taking (2) rather than (1) as basic for present purposes is not that anything is wrong with (1) as a *ceteris paribus* generalization in its own right. Rather, the point is just that (1) is too close to what we are trying to explain to be an appropriate starting point for an illuminating explanation.

Starting from (2) rather than (1), one can still explain why it is good for an agent to have true beliefs and desires for what is good for it. For if it desires that  $P$ , believes that if it does  $A$  then  $P$ , and acts so that it believes that it does  $A$ , then  $P$  if both beliefs are true, which is good for it if its desire is for what is good for it. Unfortunately, such a derivation explains much less than it appears to. For, given (2), one can show in the same way for infinitely many deviant properties  $\text{true}^*$  and  $\text{good}^*$  that the combination of  $\text{true}^*$  beliefs and desires for what is  $\text{good}^*$  for one yields (*ceteris paribus*) what is good (not just  $\text{good}^*$ ) for one.

To see this, consider an arbitrary mapping on propositions, taking the proposition that  $P$  to the proposition that  $\wedge P$ , subject to the

constraint that it commutes with logical operations, in the sense that the proposition that  $\wedge$ (not P) is the proposition that not  $\wedge$ P, the proposition that  $\wedge$ (if P then Q) is the proposition that if  $\wedge$ P then  $\wedge$ Q, and so on. In other respects, the mapping is arbitrary: for example, the proposition that  $\wedge$ (I am going north) can be the proposition that you are eating slowly.

If a proposition is just the set of possible worlds in which it is true, then we can construct such a mapping for any permutation  $\pi$  of possible worlds (a one-one mapping of the possible worlds onto the possible worlds) by stipulating that each world  $w$  belongs to the proposition that  $\wedge$ P if and only if  $\pi(w)$  belongs to the proposition that P. The mapping commutes with negation, for example, because, for any world  $w$ , the following are equivalent:  $w$  belongs to the proposition that  $\wedge$ (not P);  $\pi(w)$  belongs to the proposition that not P;  $\pi(w)$  does not belong to the proposition that P;  $w$  does not belong to the proposition that  $\wedge$ P;  $w$  belongs to the proposition that not  $\wedge$ P. For similar reasons the mapping commutes with other logical operations, such as the truth-functional conditional.

Alternatively, if propositions have quasi-syntactic structure, we can take an arbitrary permutation of their atomic constituents and extend it recursively to complex propositions in the natural way. The mapping automatically commutes with logical operations because the commutativity clauses are built into its inductive definition.

Now define “true\*” and “good\*” by these equivalences:

- (3) That P is true\* if and only if that  $\wedge$ P is true.
- (4) That P is good\* for an agent if and only if that  $\wedge$ P is good for it.

Suppose that an agent desires that P, believes that if it does A then P, and acts so that it believes that it does A. Suppose further that both beliefs are true\*. By (3), since the proposition that if it does A then P is true\*, the proposition that  $\wedge$ (if it does A then P) is true. Since the mapping commutes with logical operations, in particular with the truth-functional conditional employed (by stipulation) in (1) and (2), the proposition that  $\wedge$ (if it does A then P) is the proposition that if  $\wedge$ (it does A) then  $\wedge$ P. Thus the proposition that if  $\wedge$ (it does A) then  $\wedge$ P is true. By (3) again, since the proposition that it does A is true\*, the proposition that  $\wedge$ (it does A) is true. Since truth is closed under modus

ponens, the proposition that  $\wedge P$  is true. Suppose finally that what the agent desires is good\* for it. So that  $P$  is good\* for it; therefore, by (4), that  $\wedge P$  is good for it. In other words, something (that  $\wedge P$ ) good for the agent obtains: together, true\* belief and desire for what is good\* for one yield (*ceteris paribus*) what is good (not just good\*) for one.

From (2), we cannot conclude that the combination of true belief and desire for what is good for one is any better for one than the combination of true\* belief and desire for what is good\* for one. Yet, despite all the evolutionary pressures, we have no special tendency to believe what is true\* or to desire what is good\* for us. For example, that I am going north may be true\* if and only if you are eating slowly, and that I reach the oasis may be good\* for me if and only if it is good for me that you read your book. I have no special tendency to believe that I am going north only if you are in fact eating slowly or to desire that I reach the oasis only if it is in fact good for me that you read your book. If we start theorizing without any reason to expect a correlation between belief and truth, considerations of survival will not make the connection for us.

We can envisage schemes for interpreting creatures under which they tend to believe the true\* and desire the good\* for them, rather than to believe the true and desire the good for them. Suppose that we are trying to understand some aliens. We already have an extremely plausible interpretation  $\text{Int}$  of their beliefs and desires, under which they tend to believe the true and desire the good for them. We define a new interpretation  $\text{Int}^*$  by specifying that, under  $\text{Int}^*$ , an alien believes that  $\wedge P$  if and only if, under  $\text{Int}$ , it believes that  $P$ , and, under  $\text{Int}^*$ , it desires that  $\wedge P$  if and only if, under  $\text{Int}$ , it desires that  $P$ .<sup>5</sup> Thus  $\text{Int}^*$  ascribes a true belief just where  $\text{Int}$  ascribes a true\* belief;  $\text{Int}^*$  ascribes a desire for what is in fact good for one just where  $\text{Int}$  ascribes a desire for what is in fact good\* for one.  $\text{Int}^*$  attributes bizarre contents to the aliens: under  $\text{Int}^*$ , their beliefs about their environment have no tendency to be true, their bodily movements no tendency to bring about the satisfaction of their desires. For example,

<sup>5</sup> The definition of  $\text{Int}^*$  assumes that the proposition that  $\wedge P$  is the proposition that  $\wedge Q$  if and only if the proposition that  $P$  is the proposition that  $Q$ ; this condition is easily met.  $\text{Int}^*$  is also stipulated to ascribe to the aliens only beliefs and desires of the form that  $\wedge P$ .

under Int, an alien desires that it will be cool and believes that if it jumps into the lake then it will be cool; it jumps into the lake and will be cool. Under Int\*, it desires that  $\wedge$ (it will be cool) and believes that  $\wedge$ (if it jumps into the lake then it will be cool), in other words, that if  $\wedge$ (it jumps into the lake) then  $\wedge$ (it will be cool); it jumps into the lake and will be cool. For definiteness, let that  $\wedge$ (it will be cool) and that  $\wedge$ (it jumps into the lake) be that you were tall and that you went to bed respectively. Thus, under Int\*, the alien desires that you were tall and believes that if you went to bed then you were tall; it jumps into the lake and will be cool. Under Int, when it jumps into the lake it also believes that it jumps into the lake and that it will be cool. Thus, under Int\*, when it jumps into the lake it believes that you went to bed and that you were tall. Int\* make the aliens' mental lives formally as rational and coherent in propositional content as Int does; but Int\* radically disconnects their mental lives from what is happening around them and from what they are physically doing, whereas Int connects them in the normal way. Moreover, Int\* postulates no special mechanism to help explain the strange disconnection. Surely Int\* misinterprets the aliens. Even if such radical disconnection is not metaphysically impossible, it would occur only under highly abnormal circumstances. The nature of mental content seems to favor Int over Int\* in some constitutive way.

We could try to rule out Int\* by proposing more specific constraints on the internal interconnections of propositional attitudes for Int\* to fail. But that approach is unpromising; it misses the point of the problem. The deviant interpretation Int\* can meet even more specific constraints on the internal structure of the agent's system of propositional attitudes while still attributing mental lives radically disconnected from the environment and bodily behavior. For the mapping  $\wedge$  preserves the main structural features of propositions, and could be tailored to preserve even finer-grained structure.

It may be objected that truth\* and goodness\* are less natural properties than truth and goodness, just as grue and bleen are less natural than green and blue (Lewis 1983b). Although green will coincide with grue until some future moment, we have evolved a tendency to react differentially to green rather than to grue (even when they diverge) because green is a more natural property than grue, so a mechanism sensitive to green can develop far more easily

than a mechanism sensitive to *grue*. Evolutionary selection does not have a completely free hand; it is constrained by the available material and its causal powers. Why not explain the tendency to believe the true and desire the good for one through the combination of constraints of internal coherence such as (2) with considerations of naturalness?

A difficulty for the proposal is that *truth\** and *goodness\** need not be much more unnatural than *truth* and *goodness*. For we can define the mapping  $\wedge$  of propositions in quite natural ways, while still preserving constraints of internal coherence. For example, suppose that propositions are sets of possible worlds. Then the permutation  $\pi$  of possible worlds used to define  $\wedge$  might be a rotation of the similarity spheres of worlds about some counterfactual world. Thus each proposition that  $\wedge P$  would have the same shape in similarity space as the proposition that  $P$ , and their locations would be systematically related. Alternatively, if propositions have quasi-syntactic structure, we could replace all atomic predicative constituents of the proposition that  $P$  by their negations in constructing the proposition that  $\wedge P$ . Although such mappings may involve some loss of naturalness, it is comparatively slight. Indeed, we may even gain naturalness by selecting more natural entities than the “right” ones out of which to construct the “wrong” propositions. Yet the proposition that  $\wedge P$  will differ in truth-value from the proposition that  $P$  in very many cases; *truth\** is very poorly correlated with *truth*, and *goodness\** with *goodness*. Thus some wildly deviant interpretations  $\text{Int}^*$  are approximately as natural as or even more natural than the non-deviant interpretation  $\text{Int}$ . Moreover, the propositions we ordinarily entertain do not concern only *very* natural objects, properties and relations, for we do not ordinarily think in terms that figure in the fundamental laws of the universe. The proposition that this car is green does not cut nature at its most fundamental joints; this car is not a very natural object and greenness is not a very natural property. Nor are the properties of believing truly and desiring what is good for one very natural. At best, propositional attitude ascriptions proceed at a level of moderate naturalness. Thus the combination of constraints of internal coherence with considerations of naturalness is quite insufficient to explain why  $\text{Int}^*$  is a hopeless interpretation.

Of course, evolution *does* to some extent favor believing what is true and desiring what is good for one. But one cannot understand

why it does so simply by appeal to internal constraints and considerations of naturalness. For that understanding, one must start with a richer conception of belief and desire. More specifically, we need external constraints on the relation between mental life and the non-mental world. Much contemporary philosophy consists of attempts to provide such constraints.

### 3

Attempts to impose external constraints on the relation between mental life and the non-mental world may be roughly divided into the molecular and the holistic.<sup>6</sup> Molecularists analyze mental contents into constituents, and try to specify conditions for employing each constituent in thought. For example, a simple theory of possession conditions for concepts says that to possess the concept *mountain* one must, under optimal conditions specified without ascription of that very concept, be willing to judge *here is a mountain* if and only if a mountain is present. A simple verificationist theory of meaning states necessary and sufficient conditions for the sentence “Here is a mountain” to be canonically verified (or assertable). A simple causal theory of reference says that a thought token refers to mountains if and only if it is causally related in a specified way to mountains. And so on. More complex and sophisticated accounts can be developed in the same spirit.

If a molecularist account could be made to work, it might support many of the conclusions of this chapter. However, molecularist accounts face major obstacles. For instance, it is hard for an account that is intended to provide non-circular necessary conditions for concept possession to say anything non-trivial about what the agent does in non-optimal conditions, where ignorance and error are rife even among those who possess the concepts at issue; yet it is hard for an account to provide non-circular sufficient conditions for concept possession if it says nothing non-trivial about what the subject does in non-optimal conditions.

<sup>6</sup> The terminology of “holism” and “molecularism” is hijacked from Dummett (1975b) to make a slightly different distinction.

It is also hard to screen out the effects of the subject's background theory without circularity. As seen in earlier chapters, radical unorthodoxy is compatible with concept possession and linguistic understanding. For example, if the optimal conditions are specified without ascription of the concept *mountain*, then they can presumably be met when a revisionary metaphysician, a native English speaker with good eyesight and open eyes, dissents in good visibility from the sentence "Here is a mountain" in the middle of the Alps. The danger is that a molecularist possession condition would count her as lacking the concept *mountain*, a highly implausible result. By any reasonable standard she had the concept *mountain* before she developed her revisionary metaphysics; since she fully understood the English word "mountain," she knew that it meant *mountain*. Developing her revisionary metaphysics did not make her cease to understand the word "mountain"; she understands the word in the normal way as used by other speakers, and therefore knows that it means *mountain*; she still has the concept *mountain*. When she denies that there are mountains, she is consciously disagreeing with common sense, not talking past it. Similar problems plague verificationist theories of meaning. Not even causal theories of reference are free of such problems. Mountains may cease to cause tokenings of "mountain" in speakers with unorthodox background beliefs who continue to understand the word "mountain." Nor are causal connections always needed. Even for mountains, a community might think about them without ever having had any causal contact with them, by having causal contact with hills and envisaging mountains like hills, only bigger. As usual, attempts to preserve the necessity of the alleged condition for concept possession or linguistic understanding tend to undermine its non-circular sufficiency.

The history of molecularist programs gives little ground for optimism that such obstacles will eventually be overcome. That is not to imply that all molecularist claims are hopelessly false. Many of them seem to be true "for the most part." What is doubtful is that they can be replaced by strictly true claims within the spirit of a molecularist program.

The alternative to molecularism is holism. Although holism need not deny that thoughts have constituent structure, its constraints on thinking given thoughts apply at the level of the subject's total system of thoughts, not at the level of individual constituents; they are global

rather than local. The most salient holistic proposal is Donald Davidson's principle of charity. According to Davidson (1974: 197): "Charity is forced on us; whether we like it or not, if we want to understand others, we must count them right in most matters." He argues that methodologically good interpretation imputes agreement in the main between interpreter and interpreted; there is no obstacle in principle to a methodologically good omniscient interpreter, agreement with whom guarantees truth; since the omniscient interpreter's interpretation is by hypothesis correct, correct interpretation imputes truth in the main (1977: 200–1). Thus, by Davidson's lights, revisionary metaphysicians are bad interpreters if they interpret ordinary people as in massive error, for example over the existence of mountains. Of course, a revisionary metaphysician might claim that ordinary people do not really believe that there are mountains, but that seems to be an even worse misinterpretation. Davidson's account directly implies a tendency for beliefs to be true.

Davidson's principle of charity evokes massive disagreement. However, it is not wholly to blame for the contentious conclusions that Davidson uses it to draw. It figures in his notorious argument against the very idea of mutually incommensurable conceptual schemes, alien ways of thought or untranslatable languages (1974). But that argument also makes both the verificationist assumption that other creatures have beliefs only if we can have good evidence that they have beliefs and the constructivist assumption that we can have good evidence that they have beliefs only if we can have good evidence as to which beliefs they have. Neither assumption follows from the principle that beliefs tend to be true. Neither assumption is warranted, for we are far from omniscient interpreters (compare Nagel 1986: 93–9). The aliens may be able to interpret each other even if we cannot interpret them. More generally, Davidson's application of the methodology of radical interpretation to the philosophy of language embodies a kind of ideal verificationism, on which agents have just the intentional states that a methodologically good interpreter with unlimited access to non-intentional data would ascribe to them. However, we could, as David Lewis (1974: 110–11) recommends, treat the predicament of the radical interpreter as merely a literary device for dramatizing the question: how do the intentional states of agents supervene on the non-intentional states of the world? The sense in which that question concerns the determination of

content is metaphysical, not epistemological. In this spirit, we could consistently accept a principle of charity while allowing that alternative conceptual schemes are possible.<sup>7</sup>

If the role of the radical interpreter is inessential, so too is that of agreement between interpreter and interpreted. Truth is prior to agreement: the metaphysical version of Davidson's principle of charity requires that agents have mostly true beliefs. Other things equal, interpretation should maximize the ascribed proportion of true beliefs. That is in effect a constraint on reference for the constituents of beliefs or of the sentences that express them. Agreement is secondary; two agents with mostly true beliefs do not mostly disagree with each other, although they may have few beliefs in common, if they have different concerns, and may even tend to disagree over their limited common concerns.

Davidson's principle of charity is too loose to figure in an algorithm for reducing the intentional to the non-intentional. But present purposes do not force us to engage in the heroically ambitious quest for such a reduction. What we need are correct non-trivial principles about propositional attitudes that somehow link belief and truth, metaphysically rather than epistemologically. Such principles can fall far short of reducing the intentional to the non-intentional, even of fixing the supervenience of the former on the latter.

Even in its de-epistemologized, non-reductive version, Davidson's principle of charity remains highly contentious. Massive error seems genuinely possible for a brain envatted only months ago.<sup>8</sup> Some have responded by formulating revised principles that allow one to interpret another as in massive error when one would have been in massive error oneself in her circumstances. For example, Richard Grandy (1973: 443) proposes "as a pragmatic constraint on translation" a *principle of humanity*: "the condition that the imputed pattern of relations among beliefs, desires, and the world be as similar to our own as possible." Even if we treat the principle of humanity as a metaphysical constraint on what makes an ascription of content

<sup>7</sup> By contrast, McGinn (1986) treats radical interpretation as an epistemological problem. He explicitly allows for uninterpretable believers (367). For a recent discussion of Davidson on radical interpretation see McCulloch (2003: 94–108).

<sup>8</sup> Klein (1986) discusses of Davidson's treatment of skeptical scenarios. McCulloch (2003: 126–40) is a recent discussion of the difficulty of interpreting brains in vats.

correct, rather than an epistemological guide to plausible translation, it says nothing directly about any tendency for beliefs to be true. However, since each of our beliefs commits us to regarding it as true, and therefore as having that relation to the world, one could argue that the principle of humanity requires the beliefs of others to tend to have the same relation to the world, and therefore to be true too. Perhaps humanity implies at least a limited version of charity, although the vagueness of “similarity” between patterns of relations makes it hard to tell. But the anthropocentrism of the metaphysical principle of humanity is suspect. After all, humans are prone to peculiar logical and statistical fallacies: once we recognize a quirky design fault in ourselves, it would be perverse to prefer, on metaphysical principle, interpretations of non-human aliens that attribute the same design fault to them. Although humans are the clearest examples of rational agents with which we are familiar, we are also clear that there could be far more rational agents than we are. On their metaphysical reading, anthropocentric principles of charity implausibly imply that the very nature of content militates against the possibility of superhuman rationality.

Other principles of charity put a premium on rationality or coherence, conceived as conditions internal to the agent. But they do not explain the superiority of the sensible interpretation *Int* over the silly *Int\** above. Even those which enjoin the minimization of *inexplicable* error or ignorance rely on there being further principles, so far unspecified, for explaining error and ignorance when they are legitimately attributed: whatever those further principles are, they will do much of the work in specifying the relations between mind and world. We need to make a new start.

## 4

Suppose that Emanuel has an ill-founded faith in his ability to discern character and life-history in a face. On that basis he forms elaborate beliefs about passers-by, in which he is confident enough to bet large sums when the opportunity offers, which it rarely does. By sheer luck he has won such bets so far, which has increased his confidence in his powers, although many other beliefs he has formed in this way are in fact false. Now Emanuel sees a stranger, Celia, standing some

distance away. Looking at her face, he judges “She is F, G, H, . . .”; he ascribes a character and life-history in considerable detail. In fact, none of it fits Celia. By pure coincidence, all of it fits someone else, Elsie, whom Emanuel has never seen or heard of. Does the pronoun “she” as used by Emanuel in this context refer to Celia or to Elsie? Which of them does he use it to express beliefs about? He accepts “She is standing in front of me,” which is true if “she” refers to Celia but false if it refers to Elsie. However, he also accepts “She is F,” “She is G,” “She is H, . . .,” all of which are false if “she” refers to Celia but true if it refers to Elsie. We may assume that the latter group far outweighs the former. A principle of charity that crudely maximizes true belief or minimizes error therefore favors Elsie over Celia as the referent of the pronoun in that context. But that is a descriptive theory of reference gone mad. Emanuel has no beliefs about Elsie. He has many beliefs about Celia, most of them false. In virtue of what is Emanuel thinking about Celia rather than Elsie?

A causal theorist of reference will point out that Emanuel’s use of “she” in this context is causally related to Celia. Of course, it may be causally related to Elsie too – she may have saved Celia’s life by performing the plastic surgery on Celia’s face that helped cause Emanuel’s beliefs – but not in the right way for reference, whatever that is. In this case, the specific link is that Emanuel is perceptually attending to Celia and using “she” as a perceptual demonstrative. But to say that he is using “she” as a perceptual demonstrative is to say little more than that he is using it so as to refer to what he is perceptually attending to, and we may hope to say something more useful about what sets up this link between perception and reference. If the notion of perceptual attention is purely causal, and does not involve the notion of thinking about, in virtue of what is Emanuel thinking about that to which he has this causal relation? If, on the other hand, the notion of perceptual attention is not purely causal, and does already involve the notion of thinking about, in virtue of what is Emanuel perceptually attending to Celia? Although it is somewhat obscure just what such “in virtue of” questions are demanding, we do not simply want to meet them with silence.

A natural idea is this. The perceptual link from Celia to Emanuel matters because it is a channel for *knowledge*. If “she” refers to Celia, then, in the circumstances, Emanuel expresses knowledge when he says “She is standing in front of me,” although of course not when

he says “She is F,” “She is G,” “She is H, . . . ,” since they are false. If “she” refers to Elsie, then of course Emanuel does not express knowledge when he says “She is standing in front of me,” since it is false, but he also fails to express knowledge when he says “She is F,” “She is G,” “She is H, . . . ,” even though they are true. Emanuel is in a position to know of Celia that she is standing in front of him; he is not in a position to know of Elsie that she is F, G, H, . . . The same contrast holds, more fundamentally, at the level of thought. The assignment of Elsie as the referent in Emanuel’s beliefs gains no credit from making them true because it does not make them knowledge. The assignment of Celia wins because it does better with respect to knowledge, even though it does worse with respect to true belief.

Such examples are of course just the analogue for demonstrative pronouns of examples Kripke and Putnam used to refute descriptive cluster theories of reference for proper names and natural kind terms. In effect, such theories are special cases of a truth-maximizing principle of charity. One fundamental error in descriptive theories of reference is to try to make true belief do the work of knowledge.

As for causal theories of reference, the postulated link between knowledge and reference suggests a schematic explanation of both their successes and their failures. Roughly: a causal connection to an object (property, relation, . . .) is a channel for reference to it if and only if it is a channel for the acquisition of knowledge about the object (property, relation, . . .). Often, a causal connection is a channel for both. Equally, a non-causal connection, such as a definite description, to an object (property, relation, . . .) is a channel for reference to it if and only if it is a channel for the acquisition of knowledge about the object (property, relation, . . .). Sometimes, a non-causal connection is a channel for both. It was in any case clear that causal theories of reference and causal theories of knowledge were closely linked in their successes and failures. Both faced the problem of deviant causal chains, of specifying which causal chains carry the relevant intentional link. Both faced the problem of mathematics, which appears to exhibit both non-causal reference to abstract objects and non-causal knowledge about them.

The proposal is to replace true belief by knowledge in a principle of charity constitutive of content. But how can doing so help with the objection that massive error is possible? Presumably knowledge implies true belief. Unless the agent is inconsistent, any case of massive

error is also a case of massive ignorance. At first sight, the objection only makes the problem worse. However, it is independently obvious that our knowledge is dwarfed by our ignorance. The right charitable injunction for an assignment of reference is to maximize knowledge, not to minimize ignorance (which is always infinite).<sup>9</sup>

Suppose that under some assignment of reference a brain in a vat has mainly true beliefs about electrical impulses in the computer that controls it. If we are still disinclined to accept the assignment, a natural reason to give is that the brain is not in a position to know about the electrical impulses. If we are inclined to accept the assignment, we probably think that the brain is in a position to know about them.

Here is a simpler case. A fair coin was tossed and landed heads. The agent cannot see or otherwise know which way up it landed, but is easily convinced by what are really just his own guesses. He sincerely asserts “Toda.” Is a point in favor of interpreting “Toda” to mean “It landed heads” rather than “It landed tails” that it has him speaking and believing truly rather than falsely? Surely not. The true belief would no more be knowledge than the false belief would be. Although Davidson’s principle of charity does not imply that “Toda” cannot mean “It landed tails,” since data from other cases might outweigh the current data, it does imply that this case provides a defeasible consideration in favor of interpreting “Toda” as “It landed heads” rather than “It landed tails,” which it does not. The point extends to less irrational beliefs. If we interpret someone as judging on purely probabilistic grounds that ticket  $n$  did not win the lottery, our interpretation gains or loses no credit dependent on whether ticket  $n$  did in fact win, since either way the agent in the circumstances could not have known that it did not win.<sup>10</sup>

Is knowledge maximization in danger of absurdly imputing knowledge of quantum mechanics to Stone Age people? They were in no

<sup>9</sup> The substitution of knowledge for truth in a principle of charity is proposed in connection with a knowledge-based account of assertion by Williamson (2000a: 267).

<sup>10</sup> An interpretation on which the agent believes that ticket  $n$  did not win might do better than one on which the agent believes that ticket  $n$  won, even though neither constitutes knowledge, if the former attributes more knowledge of chances to the agent than the latter does.

position to know about quantum mechanics, so even on an interpretation on which they referred to quantum mechanical properties and relations they would not know about those properties and relations. Objective limits on what subjects are in a position to know appropriately constrain the maximization of knowledge by the assignment of reference. Unless it is raining, one does not know that it is raining. Even if it is raining, one may lack the kind of causal contact with the rain one needs in order to know that it is raining. The compositional structure of sentences and thoughts further constrains the ascription of knowledge, because the inferential processes in which subjects engage are sensitive to that structure: to interpret those processes as yielding knowledge, one must interpret them as valid inferences. Knowledge maximization need not make the ascription of knowledge come too cheap. By contrast, Davidson's principle of charity gives good marks to an interpretation for having Stone Age people assent to many truths of quantum mechanics, if it happens to fit the compositional structure of their language.

One might still fear that the knowledge maximization principle is over-charitable. Suppose, for example, that I can see only a small part of a ball, the rest of which is hidden by some obstacle. I judge of the ball "It is red." Unknown to me, the rest of the ball is green, so that the ball as a whole does not qualify as red. I falsely believe, and do not know, that the ball is red; at best I know that the visible part of the ball is red. Does knowledge maximization imply, falsely, that the visual demonstrative "it" refers to just the presently visible part of the ball rather than to the whole ball? Ask first why the visual demonstrative does not refer to the ball part. One answer is that since the ball is a more natural object than the ball part, it is a more eligible referent; I refer to the ball by default because I have done nothing special to divert reference to the ball part. Equally, then, I have failed to do the individuative work required to know anything about the ball part. By contrast, I can express some knowledge about the ball, for example, by "It is there," if "it" refers to the ball. An alternative answer is that I have positively individuated the ball, for example because my basic judgment was "That thing is red," in a thick sense of "thing" applicable to the ball but not to the ball part, from which in effect I derived "It is red" using the identity "It is that thing." But then "It is red" expresses knowledge only if "That thing is red" and "It is that thing" express knowledge and the

inference is valid, so “it” and “that thing” remain constant in reference across premises and conclusion. But if “it” refers to the ball part on both occurrences, then both premises are false, since “that thing” refers to the ball, so the conclusion fails to express knowledge. By contrast, if “it” refers anaphorically on “that thing” to the ball, “It is that thing” expresses knowledge, even though the other premise and the conclusion are false. Of course, there are further possibilities. But it is already appreciable that the holistic character of the considerations gives plenty of scope for the knowledge maximization principle to get the right answer, arguably for the right reasons.

Another doubt about knowledge maximization concerns variants of the Celia/Elsie case above in which Emanuel knows independently that Elsie is F, G, H, . . . However, he can still use “she” as a visual demonstrative to refer to Celia in judging “She is F,” “She is G,” “She is H, . . .,” thereby expressing false beliefs about Celia rather than knowledge about Elsie, because those judgments are not causally based on his independent knowledge of Elsie, and therefore fail to express that knowledge. Of course, in a further variant of the case, Emanuel makes the identity judgment “She is Elsie,” and then judges “She is F,” “She is G,” “She is H, . . .,” on the basis of inference from the identity judgment and the premises “Elsie is F,” “Elsie is G,” “Elsie is H, . . .,” so that his independent knowledge of Elsie is causally active in his reaching the conclusions. Even in that case, knowledge maximization still does not warrant assigning Elsie as the referent of the visual demonstrative “she.” If knowledge is sensitive to differences in mode of presentation, and “she” is associated with a visual mode of presentation, then the judgment “She is Elsie” does not constitute knowledge; consequently, the further judgments derived from it also fail to constitute knowledge. On the other hand, if knowledge is not sensitive to differences in mode of presentation, then assigning Elsie as the referent of “she” merely makes the judgments “She is F,” “She is G,” “She is H, . . .,” express the same knowledge as “Elsie is F,” “Elsie is G,” “Elsie is H, . . .,” already express; no knowledge is gained. Moreover, that assignment also makes judgments such as “She is standing in front of me” fail to constitute knowledge, whereas they do constitute knowledge on the assignment of Celia as the referent of “she.” Hence the correct assignment (Celia) involves the ascription of more knowledge than the incorrect one

(Elsie) does. Thus knowledge maximization is consistent with a correct interpretation of such cases.

Perhaps the underlying worries about knowledge maximization can be captured in a more abstract form. Knowing is itself an intentional state. As already emphasized, the present aim is not to reduce the intentional to the non-intentional. But to explain reference by appeal to the intentional features of knowledge states – which objects, properties, and relations they are about – is in effect to explain reference in terms of itself. In order to avoid such trivialization, we must avoid helping ourselves to those intentional features, and instead concentrate on the imputed reliability of the subject (in some appropriate sense) under various assignments of reference (where such assignments assign reference across many possible worlds). But if that is what we have to maximize, surely the winner is likely to be some artificial cooked-up assignment quite different from what is pretheoretically correct. For example, a highly context-sensitive assignment may make the Stone Age people reliable about matters of quantum mechanics. Similarly, some assignment will make the victim of a skeptical scenario come out thinking reliably about their own brain states rather than unreliably about the wider world. And so on. How can knowledge maximization avoid such false consequences without collapsing into triviality?

We take such assignments of reference to be incorrect because we take them to be gerrymandered, unnatural, insensitive to the underlying similarities and differences, not cutting at the joints. The corresponding ascriptions of knowledge make it an equally artificial attitude. In response to such examples, we should therefore insist that the relation to be maximized is a natural one: doubtless not a *perfectly* natural one, for the most basic structure of the world is not mental, but natural by the standards of mentality. Such a bias towards naturalness in the objects of reference has independent support (Lewis 1983b, Weatherson 2003, Hawthorne 2006: 53–69). Here it is extended to the relation of reference itself, by inheritance from the relation of knowledge. It holds the anti-skeptical effect of knowledge maximization within reasonable limits.

The more abundant ontology is, the more objects, properties, and relations there are, the more scope there is for an assignment of reference under which we know. Conversely, the sparser ontology is, the fewer objects, properties, and relations there are, the greater the

danger that we do not know under any assignment. But the correlation is imperfect, for a sparse ontology sometimes facilitates knowledge by reducing the number of wrong answers clustered around the right one and hard to distinguish from it. Knowledge maximization tilts the playing field in our favor without guaranteeing us victory.

Is it surprising that reference maximizes knowledge? Reference concerns what mental states and acts are about. Knowledge is one mental state among many. Why should it play a privileged role in determining what all of them are about? One answer is that knowledge is not just one mental state among many. A creature that is not aware of anything at all has no mental life. It lacks genuine intelligence. Although intelligent life does not consist solely of awareness, it is intelligent only because appropriately related to awareness of something. But to be aware is to know: one is aware *that* P if and only if one knows that P, and one could hardly be aware of anything without some capacity to know *that* something is the case. Intelligent life is life appropriately related to intelligent action, and intelligent action is action appropriately related to knowledge. In a paradigm of intelligent action, given a desire that P, one knowingly does A, knowing that if one does A then P. One can believe that one does A and that if one does A then P, even truly, without knowing, but the action is defective in such cases; they are to be understood in relation to non-defective cases. The function of intelligent action involves the application of knowledge to realize the agent's ends. In unfavorable circumstances, only mere beliefs are available, and intentional action does not function properly, although with good luck it may still achieve the desired end, just as other defective processes sometimes issue in the intended product.<sup>11</sup>

<sup>11</sup> Williamson (2000a) has more on the associated conception of mind and knowledge. The idea that all thinking qualifies as such by being appropriately related to knowing was advocated by another Wykeham Professor of Logic, John Cook Wilson (1926, vol. I: 35–40, also for the view that knowledge is indefinable). He defends a neo-Aristotelian version of common sense realism on which ordinary language has a central role in metaphysics. Of the “examination of the meaning of grammatical forms” and the consideration of “certain distinctions of the kind called metaphysical” he says “The two investigations are necessarily connected with one another; for since the sentence or statement describes the nature of objects and not any attitude of ours to the objects described, in the way of apprehension or opinion, its meaning is wholly objective, in the sense that we have already given to objective. That is, it is about

When conditions are unfavorable, the agent is in no position to know anything much, just as a victim of total paralysis may be in no position to do anything much. Intentional action may be limited to pursuing a line of thought. For a brain in a vat, both knowledge and action may shrink to the internal: but that pathological case does not reveal their underlying nature, for it does not show them to be equally shrunken in more normal cases. Rather, the pathological cases are parasitic on the normal ones.

Given the central role of knowledge in intelligent life, the intimate relation between knowledge and reference is hardly surprising. Reference maximizes knowledge because its role is to serve knowledge, not to impose any independent limitation on it. Although maximizing knowledge is not equivalent to maximizing true belief, the nature of reference grounds a general, highly defeasible tendency for beliefs to constitute knowledge, and therefore to be true.

## 5

On a more internalist proposal, the nature of reference is to maximize justified belief rather than knowledge, where justified beliefs can be false; charity is often presented as a principle of rationality maximization. But such internalism makes the bearing of reference on justification obscure. Suppose that I have a few factual memories of a brief acquaintance, which I express using the pronoun “he.” The assignment of one reference rather than another to “he” seems to make no difference to the internalist justification of my memory beliefs; it makes an obvious difference to whether they constitute knowledge. Similarly, internalist considerations of justified belief are much less likely than externalist considerations of knowledge to explain why the silly interpretation *Int\** in Section 2 is worse than the sensible

---

something apprehended, in the case of knowledge for instance, and not about our apprehension of it” (1926, vol. I: 149). In respect of the fundamental role assigned to knowing, both Williamson (2000a) and the present book belong to a tradition that runs from Cook Wilson to Prichard and others, then to J.L. Austin and later to John McDowell; see Marion (2000). That there are also very significant differences between these philosophers hardly needs saying.

connection Int, for the permutation of contents preserves internal coherence but not knowledge.

On an uneasy compromise, what matters for reference is neither knowledge nor internalist justification but an intermediate standard of non-factive externalist justification. That still gets it wrong, because the failure of a brain in a vat to refer to a new object in its external environment is far better explained by its incapacity for knowledge of it than by its incapacity for justified (and perhaps true) beliefs about it, for on the supposition that it has beliefs about the object there need be no further obstacle to classifying them as justified in the relevant sense. By contrast, the full-blooded external involvement of knowledge exactly suits it to constrain reference.

Can the semantic significance of knowledge be understood within Davidson's framework? He tries to recover a plausible epistemology by extracting epistemological consequences from his principle of charity by appeal to the immunity from massive error that it is supposed to grant. That immunity is holistic: it is consistent with the falsity of almost any given one of our beliefs, given enough compensating truth elsewhere in the system. For example, my belief that I have hands enjoys no immunity from error. The supposed general immunity from massive error does not explain how I know that I have hands: likewise for most of what we ordinarily take ourselves to know. Davidson adds an appeal to causal constraints on reference in simple cases, but formulates the constraints too crudely to permit any straightforward connection with knowledge (Davidson 1991: 196–7). Even if my belief that P is caused by what it is about, I may fail to know that P because the causal chain is somehow deviant. When Davidson tries to explain how his principle of charity yields knowledge, he appears to rely on something like the pre-Gettier assumption that justified true belief is knowledge.<sup>12</sup>

<sup>12</sup> “There is at least a presumption that we are right about the contents of our own minds; so in the cases where we are right, we have knowledge” (Davidson 1991: 194); “Anyone who accepts perceptual externalism knows he cannot be systematically deceived about whether there are such things as cows, people, water, stars, and chewing gum. Knowing why this is the case, he must recognize situations in which he is justified in believing he is seeing water or a cow. In those cases where he is right, he knows he is seeing water or a cow” (Davidson 1991: 201). See also Davidson (1983).

A subtler attempt to extract knowledge from Davidson's principle of charity exploits beliefs that one knows. Very often, when one believes that  $P$ , one also believes that one knows that  $P$ .<sup>13</sup> If one believes truly that one knows that  $P$ , then one does know that  $P$ . Does maximizing true belief therefore indirectly maximize knowledge too? The detour through second-order belief is unpromising. First, it depends on the assumption that the relevant agents are to be interpreted as believing that they know. Of course, *we* often believe that we know; for that matter, we often know. But the aim was to derive the conclusion that agents in general often know from a truth-maximizing principle of charity; that agents in general often believe that they know has not been derived from such a principle. Second, even granted that agents believe that they know, Davidson's principle attributes no special status to beliefs of that form; an interpretation might sacrifice them all as false and still maximize true belief overall by making enough other beliefs true. Third, the account does not generate attributions of knowledge to simple creatures who lack the concept of knowledge and therefore cannot believe that they know; surely they can have knowledge without having the concept of knowledge.<sup>14</sup> Truth maximization lacks most of the epistemological rewards of knowledge maximization.

Quine endorses as a canon of translation the epistemological-sounding maxim "Save the obvious" (1970: 82; compare 1960: 59): do not interpret the natives as dissenting from obvious truths. On that

<sup>13</sup> The principle cannot be exceptionless, otherwise having any belief involves having infinitely many beliefs of increasing complexity.

<sup>14</sup> Davidson might have denied that one can have knowledge without the concept of knowledge, for he denies that one can have beliefs without the concept of belief: "Someone cannot have a belief unless he understands the possibility of being mistaken, and this requires grasping the contrast between truth and error – true belief and false belief" (1975: 170). Whether or not he would extend it to knowledge, Davidson's argument is unconvincing, for it conflates *de re* and *de dicto* readings. Grant for the sake of argument that, to believe that  $P$ , one must grasp the contrast between the state of affairs that  $P$ , which is in fact the condition for the belief to be true, and the state of affairs that not  $P$ , which is in fact the condition for the belief to be false (the *de re* reading). Even so, Davidson does not explain why one must grasp it as the contrast between the condition for the belief to be true and the condition for it to be false (the *de dicto* reading), which is what he needs. Thus he leaves it obscure why a creature with the concept of negation could not have a belief without the concept of belief.

basis he argues that apparent deviations in logic are mere artifacts of bad translation. Although this appears to invoke a knowledge-related standard of charity, like the principle of knowledge maximization, Quine insists on interpreting “obvious” behavioristically rather than epistemologically.<sup>15</sup> His intended maxim is that translation should preserve general assent. Without further argument, we cannot conclude that sentences that enjoy general assent are true, for we can assume neither that every sentence to which speakers of another language assent can be translated into English nor that every sentence to which speakers of English assent is true – naturally, it is harder for us, as speakers of English, to produce a counterexample. Like Grandy’s principle of humanity, Quine’s maxim on its behavioral reading tends to project our design faults onto others. For example, it discourages us from translating a sentence to which the natives universally assent by a simple logical truth from which many speakers of English dissent through intellectual confusion. On an epistemological reading of “obvious,” the maxim is not vulnerable to that criticism, for confused speakers can dissent from what is obvious.

We do better to start with the notion of knowledge in the explanatory order.

## 6

A picture of the mind has been sketched, with the broadest strokes, on which the nature of reference nudges belief towards the status of knowledge, and therefore of truth. That helps put the burden of proof on judgment skeptics to argue that their radical scenarios deserve to be taken more seriously than do the radical scenarios for skepticism about perception. Although we can allow that scenarios of both sorts are metaphysically possible, much more than that is needed to justify serious doubt. The burden of proof on the judgment skeptic is particu-

<sup>15</sup> “I must stress that I am using the word ‘obvious’ in an ordinary behavioral sense, with no epistemological overtones. When I call ‘1 + 1 = 2’ obvious to a community I mean only that everyone, nearly enough, will unhesitatingly assent to it, for whatever reason; and when I call ‘It is raining’ obvious in particular circumstances I mean that everyone will assent to it in those circumstances” (Quine 1970: 82).

larly heavy when the proposed scenarios make vast ranges of common beliefs false or at least not knowledge, as many of them do.<sup>16</sup>

A judgment skeptic might respond: “Granted, when we believe  $p$ , we often – but not always – know  $p$ . That we believe  $p$  should therefore be treated as good but defeasible evidence for  $p$ . It is just one more part of the total body of evidence on which philosophical theories should be evaluated.” This response depends on the fallacy, diagnosed in the previous chapter, of psychologizing evidence. It perversely ignores the evidential role of  $p$  itself, as opposed to that of the fact that we believe  $p$ . After all, if we do know  $p$ , would it not be negligent not to use that knowledge in evaluating a philosophical theory to which it is relevant? Philosophy is hard enough already: why make it even more difficult by forbidding ourselves to bring some of our knowledge to bear? You are not obliged to fight with one arm tied behind your back.<sup>17</sup>

The judgment skeptic might reply that, if we know  $p$  without knowing that we know  $p$ , the knowledge does not really help. But that response is doubly inadequate. First, it gives no more reason to deny that we know that we know  $p$  than to deny that we know  $p$  in the relevant cases. Although we cannot expect to have infinitely many iterations of knowledge, for more than computational reasons (Williamson 2000a: 114–34), that general point merely shows that we must sometimes simply apply our knowledge, without first checking whether we know, for otherwise we get stuck in an infinite regress of checks. That is the second problem for the judgment skeptic’s envisaged reply. It gave us no evidence that we are entitled to rely on the premise  $p$  in philosophical discussion only if we know that we know  $p$ .

When we know, there is something non-trivial to be said about how we know. But we may know  $p$ , and even know that we know  $p$ , without knowing how we know  $p$ . For instance, we may know that we know the truth of some logical or mathematical axioms without knowing how we know their truth. Similarly, the epistemic role of elegance and simplicity in theoretical physics seems as indis-

<sup>16</sup> The case of folk physics does not constitute a straightforward skeptical scenario, for folk physics plays a role in generating much knowledge of particular facts about our environment.

<sup>17</sup> See Williamson (2000a: 184–208) for defense and development of the conception of our total evidence as everything that we know.

pensable as it is hard to explain. But for many philosophically contentious facts, the question “How do you know?” is not unusually puzzling. There is no distinctive mystery as to how we know that there are mountains in Switzerland. We can explain how we know, typically by describing the process by which we acquired the knowledge, without having to convince the skeptic who doubts that we know.

Even those who know  $p$  can sometimes be too dogmatic about  $p$  in this sense: their summary dismissal of objections to  $p$  manifests general cognitive dispositions whose overall tendency is to limit their knowledge and increase their error, by preventing them from learning from experience or criticism.<sup>18</sup> But that does not show that they acted wrongly in treating  $p$  as evidence in this particular case. There will always be cases in which bad dispositions produce right actions and good dispositions produce wrong ones; since philosophers question fundamental assumptions, they are particularly liable to get themselves into such cases.

The knowledge maximization principle is not itself intended as an answer to the question “How do you know?” The knowledge maximized may have been acquired by quite familiar means of perception, memory, testimony, inference, and imagination. The proper response to judgment skepticism is not to postulate a separate means to knowledge to underpin all the others but rather to challenge the skeptical idea that they need such underpinning. The supposed function of the underpinning would be to rule out the scenarios that motivate judgment skepticism. But a good answer to the question “How do you know  $p$ ?” need not specifically address far-fetched skeptical scenarios for  $p$ , since knowing  $p$  does not require specifically addressing them. Knowledge maximization is a factor, typically unnoticed by judgment skeptics, that makes their scenarios more far-fetched than they realize.

More naturalistically inclined judgment skeptics try to induce a crisis of confidence in present common sense by pointing towards a present or future scientific outlook that stands to present common sense as the latter stands to a Stone Age outlook. But the analogy rebounds against judgment skepticism. For although it is plausible

<sup>18</sup> One can know  $p$  and acquire counter-evidence to  $p$  that is significant, but not significant enough to make one cease to know  $p$ .

that Stone Age people had many false beliefs about the general nature of the world, it is at least as plausible that they had significant knowledge of their local environment. Knowledge maximization implies that our ancestors had some primitive knowledge as soon as they had some primitive beliefs; it is not as though archaeology suggests otherwise. Again, if it is plausible that some non-human animals have primitive beliefs, it is equally plausible that they have some primitive knowledge.

Consider this analogue for observational evidence of the judgment skeptic's response to knowledge maximization: "Granted, when we have a perceptual belief in  $p$ , we often – but not always – know  $p$ . That we have a perceptual belief in  $p$  should therefore be treated as good but defeasible evidence for  $p$ . It is just one more part of the total body of evidence on which scientific theories should be evaluated." What this response perversely ignores is the evidential role of  $p$  itself, as opposed to that of the fact that we have a perceptual belief in  $p$ . After all, if we do know  $p$ , would it not be negligent not to use that knowledge in evaluating a scientific theory to which it is relevant? It would not advance science to insist that scientists' evidence cannot include the fact that 19 out of 20 rats fed the substance died within 24 hours, but only the fact that the scientist had the perceptual belief that 19 out of 20 rats fed the substance died (only the former fact leads itself to statistical analysis). Such claims about past beliefs are not peculiarly foundational. Indeed, they are less amenable to public checking by the scientific community than are claims about the actual outcomes of experiments. Of course, it may later turn out that a disgruntled lab technician fed the rats the wrong substance, but the proper response to such remote possibilities is to backtrack if one of them is found to obtain, not to make a futile attempt in advance to identify evidence for which backtracking will never be required in even the remotest eventualities.

In philosophy as in natural science, our evidence consists of ordinary human knowledge. We have no general guarantee that we know everything we think we know. Our evidence is more contested in philosophy than in natural science. The philosopher's predicament is somewhat like that which would face natural scientists if accusations of falsified evidence were vastly more common in science than they currently are. Whatever the discipline, when someone disputes the evidence, it is often better to look for common ground on which

to pursue the argument than to ride roughshod over the objections. For that temporary purpose, we may refrain from treating the disputed evidence as evidence; that does not entail that it should never have been treated as evidence in the first place. Moreover, as we have seen, the search for common ground can be taken too far, especially with a reckless opponent who does not scruple to challenge any inconvenient evidence. An indiscriminate skeptic can challenge whatever we offer as evidence, by always demanding a proof; that should not drive us to suspend all our evidence. At some point we are entitled to hold on to what we know, and apply it.

Our evidence in philosophy consists of a miscellaneous mass of knowledge, expressed in terms of all kinds, some from ordinary language, some from the theoretical vocabulary of various disciplines. Some of it consists of knowledge about our own mental states; most of it does not. Whatever we know is legitimate evidence. Inevitably, we make mistakes, treating as known what is unknown, or as unknown what is known. The principle of knowledge maximization helps our practice survive our critical reflection, by reassuring us that knowing is a natural state for believers, not an anomalous achievement. In general, our practice makes sense, which of course does not excuse us from meeting particular challenges on their merits. This messy epistemological predicament in which philosophers find themselves is not deeply different from the messy epistemological predicament of all human inquiry.

# *Afterword*

## Must Do Better

---

Imagine a philosophy conference in Presocratic Greece. The hot question is: what are things made of? Followers of Thales say that everything is made of water, followers of Anaximenes that everything is made of air, and followers of Heraclitus that everything is made of fire. Nobody is quite clear what these claims mean; some question whether the founders of the respective schools ever made them. But among the groupies there is a buzz about all the recent exciting progress. The mockers and doubters make plenty of noise too. They point out that no resolution of the dispute between the schools is in sight. They diagnose Thales, Anaximenes, and Heraclitus as suffering from a tendency to over-generalize. We can intelligibly ask what bread is made of, or what houses are made of, but to ask what *things in general* are made of is senseless, some suggest, because the question is posed without any conception of how to verify an answer; language has gone on holiday. Paleo-pragmatists invite everyone to relax, forget their futile pseudo-inquiries, and do something useful instead.

The mockers and doubters had it easy, but we know now that in at least one important respect they were wrong. With however much confusion, Thales and the rest were asking one of the best questions ever to have been asked, a question that has painfully led to much of modern science. To have abandoned it two and a half thousand years ago on grounds of its conceptual incoherence or whatever would have been a feeble and unnecessary surrender to despair, philistinism, cowardice, or indolence. Nevertheless, it is equally clear that the methods of investigation used by the Presocratics were utterly inadequate to their ambitions. If an intellectual tradition applied just those methods to those questions for two and a half millennia, which

is far from unimaginable, it might well be very little the wiser at the end. Much of the progress made since the Presocratics consists in the development of good methods for bringing evidence to bear on questions that, when first asked, appear hopelessly elusive or naïve. Typically, of course, making progress also involves refining and clarifying the initial question: but the relevant refinements and clarifications cannot all be foreseen at the beginning. They emerge in the process of attempting to answer the original rough question, and would not emerge otherwise.

The Presocratics were forerunners of both modern philosophy and modern natural science; they did not distinguish natural science from philosophy. For positivists, the moral of the story is that natural science had to be separated from philosophy, and marked out as the field for observation, measurement, and experiment, before it could make serious progress. There is doubtless something right about that moral, although as it stands it hardly does justice to the significance of armchair methods in natural science, such as the use of mathematics and of thought experiments, for example by Galileo and Einstein. Moreover, the positivist moral misses a deeper methodological point. The case of the Presocratics shows that one cannot always tell in advance which questions will be fruitful to pursue. Even if a community starts with no remotely adequate idea of how to go about answering a question, it does not follow that the question is meaningless or not worth addressing. That goes for the questions we now classify as philosophical as much as it does for those we now classify as empirical or natural-scientific.

The opponents of systematic philosophical theorizing might reply that they are not judging philosophical questions in advance; they are judging them after two and a half millennia of futile attempts to answer them. Of course, it is an important issue how similar our philosophical questions are to those of ancient Greece, or even to those of Enlightenment Europe. Nevertheless, philosophy has been going too long as an intellectual tradition separate from natural science (although sometimes interacting with it) for the question “How much progress has it made?” to be simply dismissed as premature.

We should not be too pessimistic about the answer, at least concerning the broad, heterogeneous intellectual tradition we conveniently label “analytic philosophy.” In many areas of philosophy,

we know much more in 2007 than was known in 1957; much more was known in 1957 than in 1907; much more was known in 1907 than was known in 1857. As in natural science, something can be collectively known in a community even if it is occasionally denied by eccentric members of that community. Although fundamental disagreement is conspicuous in most areas of philosophy, the best theories in a given area are in most cases far better developed in 2007 than the best theories in that area were in 1957, and so on. Much of the knowledge is fairly specific in content. For example, we know far more about possibility and necessity than was known before the development of modern modal logic and associated work in philosophy. It is widely known in 2007 and was not widely known in 1957 that contingency is not equivalent to *a posteriori*, and that claims of contingent or temporary identity involve the rejection of standard logical laws. The principle that every truth is possibly necessary can now be shown to entail that every truth is necessary by a chain of elementary inferences in a perspicuous notation unavailable to Hegel (every instance of the schema  $A \rightarrow \Box A$  is derivable from instances of the schema  $A \rightarrow \Diamond \Box A$  in the weak modal system T). We know much about the costs and benefits of analyzing possibility and necessity in terms of possible worlds, even if we do not yet know whether such an analysis is correct.<sup>1</sup>

Another example: Far more is known in 2007 about truth than was known in 1957, as a result of technical work by philosophical and mathematical logicians such as Saul Kripke, Solomon Feferman, Anil Gupta, Vann McGee, Volker Halbach, and many others on how close a predicate in a language can come to satisfying a full disquotational schema for that very language without incurring semantic

<sup>1</sup> This guarded optimism about philosophical progress is consistent with the pessimism in Williamson (2000a) about the prospects for the post-Gettier program of analyzing the concept of knowledge and similar programs of analyzing other philosophically significant concepts. Such programs did make progress in clarifying the relations between the concepts under study (and between the things to which those concepts refer). What they failed to make plausible was that the eventual outcome of such progress would be anything like an analysis in the intended sense (necessary and sufficient conditions stated in non-circular terms, perhaps meeting further conditions). Take any concept that is indefinable in the relevant sense: the vain program of analyzing it in terms of more basic concepts, if conducted by able and honest people over several decades, would lead to some progress of this kind.

paradoxes. Their results have significant and complex implications, not yet fully absorbed, for current debates concerning deflationism and minimalism about truth (see Halbach (2001) for a recent example). One clear lesson is that claims about truth need to be formulated with extreme precision, not out of knee-jerk pedantry but because in practice correct general claims about truth often turn out to differ so subtly from provably incorrect claims that arguing in impressionistic terms is a hopelessly unreliable method. Unfortunately, much philosophical discussion of truth is still conducted in a programmatic, vague, and technically uninformed spirit whose products inspire little confidence.

In 1957, Michael Dummett was about to open his campaign to put the debate between realism and anti-realism, as he conceived it, at the centre of philosophy. The campaign had a strong methodological component. Intractable metaphysical disputes (for example, about time) were to be resolved by being reduced to questions in the philosophy of language about the proper form for a semantic theory of the relevant expressions (for example, tense markers). The realist's semantic theory would identify the meaning of an expression with its contribution to the truth conditions of declarative sentences in which it occurred. The anti-realist's semantic theory would identify the meaning with the expression's contribution to the assertability conditions of those sentences. Instead of shouting slogans at each other, Dummett's realist and anti-realist would busy themselves in developing systematic compositional semantic theories of the appropriate type, which could then be judged and compared by something like scientific standards. But that is not what happened.

True, over recent decades truth-conditional semantics for natural languages has developed out of philosophical logic and the philosophy of language into a flourishing branch of empirical linguistics. Frege already had the fundamental conception of compositional truth-conditional semantics, in which expressions refer to items in the mostly non-linguistic world, the reference of a complex expression is a function of the reference of its constituents, and the reference of a sentence determines its truth value. But Frege was more concerned to apply that conception to ideally precise and perspicuous artificial languages than to messy natural ones. The systematic application of compositional truth-conditional semantics to natural languages goes back to Richard Montague (under the influence of

Carnap) in its intensional form and has been mediated in linguistics by Barbara Partee and others. In its extensional form, it goes back to Donald Davidson (under the influence of Tarski) and has been mediated in linguistics by Jim Higginbotham and others. Needless to say, that crude schema does no justice to the richness of recent work and the variety of contributors to it (in both departments of philosophy and departments of linguistics), which one can check by looking at any decent handbook of contemporary semantic theory as a branch of linguistics. Surprisingly, however, most participants in the Dummett-inspired debates between realism and anti-realism have shown little interest in the success of truth-conditional semantics, judged as a branch of empirical linguistics. Instead, they have tended to concentrate on Dummett's demand for "non-circular" explanations of what understanding a sentence with a given truth condition "consists in," when the speaker cannot verify or falsify that condition. That demand is motivated more by preconceived philosophical reductionism than by the actual needs of empirical linguistics. Thus the construction and assessment of specific truth-conditional semantic theories has almost disappeared from sight in the debate on realism and anti-realism.

As for assertability-conditional semantics, it began with one more or less working paradigm: Heyting's intuitionistic account of the compositional semantics of mathematical language in terms of the condition for something to be a proof of a given sentence. The obvious and crucial challenge was to generalize that account to empirical language: as a first step, to develop a working assertability-conditional semantics for a toy model of some small fragment of empirical language. But that challenge was shirked. Anti-realists preferred to polish their formulations of the grand program rather than getting down to the hard and perhaps disappointing task of trying to carry it out in practice. The suggestion that the program's almost total lack of empirical success in the semantics of natural languages might constitute some evidence that it is mistaken in principle would be dismissed as crass.

Some participants in the debate denied any need for anti-realists to develop their own semantic theories of a distinctive form. For, it was proposed, anti-realists could take over truth-conditional semantic theories by interpreting "true" to mean assertable or verifiable at the limit of inquiry, or some such epistemic account of truth (Wright

1993: 403–25). But that proposal is quite contrary to Dummett's original arguments. For they require the key semantic concept in the anti-realistic semantics, the concept in terms of which the recursive compositional clauses for atomic expressions are stated, to be decidable, in the sense that the speaker is always in a position to know whether it applies in a given case. That is what allows anti-realists to claim that, unlike realists, they can give a non-circular account of what understanding a sentence consists in: a disposition to assert it when and only when its assertability condition obtains. But it is supposed to be common ground between realists and anti-realists that truth is not always decidable. A speaker may understand a sentence without being in a position either to recognize it as true or to recognize it as not true. I can understand the sentence “There was once life on Mars,” even though I have neither warrant to assert “There was once life on Mars” nor warrant to assert “There was never life on Mars.” The point is particularly clear in the intuitionistic semantics for mathematical language. The key concept in the compositional semantics is the concept *p is a proof of s*, which is decidable on the intuitionistic view because to understand a sentence is to associate it with an effective procedure for recognizing whether any given putative proof is a proof (in some canonical sense) of it. By contrast, what serves as the intuitionistic concept of truth is not the dyadic concept *p is a proof of s* nor even the monadic concept *s has been proved* but the monadic concept *s has a proof* or *s is provable*. According to intuitionists, we understand many mathematical sentences (such as “There are seven consecutive 7s in the decimal expansion of  $\pi$ ”) without having a procedure for recognizing whether they are provable. We understand them because we can recognize of any given putative proof, once presented to us, whether it is indeed a proof of them. Nor can we replace “true” in a truth-conditional semantics by “has been proved” (treated as decidable), because that would reduce the semantic clause for negation (that the negation of a sentence *s* is true if and only if *s* is not true) to the claim that the negation of *s* has been proved if and only if *s* has not been proved, which is uncontroversially false whenever *s* has not yet been decided.

Dummett's requirement that assertability be decidable forces assertability-conditional semantics to take a radically different form from that of truth-conditional semantics. Within this tradition, anti-

realists have simply failed to develop natural language semantics in that form, or even to provide serious evidence that they could so develop it if they wanted to.<sup>2</sup> They proceed as if Imre Lakatos had never promulgated the concept of a degenerating research program.

Dummett's posing of the issue between realism and anti-realism provides a case study of an occasion when the philosophical community was offered a new way of gaining theoretical control over notoriously elusive issues, through the development of systematic semantic theories. The community spurned the opportunity, if that is what it was. Those who discussed realism and anti-realism on Dummett's terms tended to concentrate on the most programmatic issues, which they debated with no more clarity or conclusiveness than was to be found in the traditional metaphysical reasoning that Dummett intended to supersede. The actual success or lack of it in applying the rival semantic programs to specific fragments of natural language was largely ignored. Far from serving as a beacon for a new methodology, the debate between realism and anti-realism has become notorious in the rest of philosophy for its obscurity, convolution, and lack of progress.

Of course, one may reject Dummett's attempted reduction of issues in metaphysics to issues in the philosophy of language. As seen in earlier chapters, not all philosophical questions are really questions about language or thought. However, as we also saw, that a question is non-semantic does not imply that semantics imposes no useful constraints on the process of answering it. To reach philosophical conclusions one must reason, usually in areas where it is very hard to distinguish valid from invalid reasoning. To make that distinction reliably, one must often attend carefully to the semantic form of the premises, the conclusion, and the intermediate steps. That requires implicit semantic beliefs about the crucial words and constructions. Sometimes, those beliefs must be tested by explicit semantic theorizing. Philosophers who refuse to bother about semantics, on the grounds that they want to study the non-linguistic world, not our

<sup>2</sup> Perhaps some work in contemporary formal semantics can be interpreted as assertability conditional rather than truth conditional in spirit: for instance, probability semantics for conditionals and other constructions, some forms of speech act theory, some theories of dynamic semantics. It is doubtful that much of this work conforms to Dummett's anti-realist constraints, or even makes a serious attempt to do so.

talk about that world, resemble scientists who refuse to bother about the theory of their instruments, on the grounds that they want to study the world, not our observation of it. Such an attitude may be good enough for amateurs; applied to more advanced inquiries, it produces crude errors. Those metaphysicians who ignore language in order not to project it onto the world are the very ones most likely to fall into just that fallacy, because their carelessness of the structure of the language in which they reason makes them insensitive to subtle differences between valid and invalid reasoning.

Explicit compositional semantic theories for reasonable fragments of particular natural languages also have the great methodological advantage of being comparatively easy to test in comparatively uncontentious ways, because they make specific predictions about the truth conditions (or assertability conditions) of infinitely many ordinary unphilosophical sentences. The attempt to provide a semantic theory that coheres with a given metaphysical claim can therefore constitute a searching test of the latter claim, even though semantics and metaphysics have different objects.

Discipline from semantics is only one kind of philosophical discipline. It is insufficient by itself for the conduct of a philosophical inquiry, and may sometimes fail to be useful, when the semantic forms of the relevant linguistic constructions are simple and obvious. But when philosophy is not disciplined by semantics, it must be disciplined by something else: syntax, logic, common sense, imaginary examples, the findings of other disciplines (mathematics, physics, biology, psychology, history, . . .) or the aesthetic evaluation of theories (elegance, simplicity, . . .). Indeed, philosophy subject to only one of those disciplines is liable to become severely distorted: several are needed simultaneously. To be “disciplined” by X here is not simply to pay lip-service to X; it is to make a systematic conscious effort to conform to the deliverances of X, where such conformity is at least somewhat easier to recognize than is the answer to the original philosophical question. Of course, each form of philosophical discipline is itself contested by some philosophers. But that is no reason to produce work that is not properly disciplined by anything. It may be a reason to welcome methodological diversity in philosophy: if different groups in philosophy give different relative weights to various sources of discipline, we can compare the long-run results of the rival ways of working. Tightly constrained work has the merit that even those

who reject the constraints can agree that it demonstrates their consequences.

Much contemporary analytic philosophy seems to be written in the tacit hope of discursively muddling through, uncontrolled by any clear methodological constraints. That may be enough for easy questions, if there are any in philosophy; it is manifestly inadequate for resolving the hard questions with which most philosophers like to engage. All too often it produces only eddies in academic fashion, without any advance in our understanding of the subject matter. Although we can make progress in philosophy, we cannot expect to do so when we are not working at the highest available level of intellectual discipline. That level is not achieved by effortless superiority. It requires a conscious collective effort.

We who classify ourselves as “analytic” philosophers tend to fall into the assumption that our allegiance automatically grants us methodological virtue. According to the crude stereotypes, analytic philosophers use arguments while “continental” philosophers do not. But within the analytic tradition many philosophers use arguments only to the extent that most “continental” philosophers do: some kind of inferential movement is observable, but it lacks the clear articulation into premises and conclusion and the explicitness about the form of the inference that much good philosophy achieves. Again according to the stereotypes, analytic philosophers write clearly while “continental” philosophers do not. But much work within the analytic tradition is obscure even when it is written in everyday words, short sentences and a relaxed, open-air spirit, because the structure of its claims is fudged where it really matters.

If the high standards that make philosophy worth doing are often absent even in analytic philosophy, that is not because they are a natural endowment found only in a brilliant elite. Even if Frege’s exceptional clarity and rigor required innate genius – although they undoubtedly also owed something to the German mathematical tradition within which he was educated – after his example they can now be effectively taught. Some graduate schools communicate something like his standards, others notably fail to do so.

Of course, we are often unable to answer an important philosophical question by rigorous argument, or even to formulate the question clearly. High standards then demand not that we should ignore the question, otherwise little progress would be made, but that we should

be open and explicit about the unclarity of the question and the inconclusiveness of our attempts to answer it, and our dissatisfaction with both should motivate attempts to improve our methods. Moreover, it must be sensible for the bulk of our research effort to be concentrated in areas where our current methods make progress more likely.

We may hope that in the long term philosophy will develop new and more decisive methods to answer its questions, as unimaginable to us as our methods were to the Presocratics. Indeed, the development of such methods is one of the central challenges facing systematic philosophy. Paul Grice once wrote “By and large the greatest philosophers have been the greatest, and the most self-conscious, methodologists; indeed, I am tempted to regard this fact as primarily accounting for their greatness as philosophers” (Grice 1986: 66). Nevertheless, we must assume, in the short term philosophy will have to make do with something like currently available methods. But that is no reason to continue doing it in a methodologically unreflective way. A profession of very variable standards can help the higher to spread at the expense of the lower, by conscious collective attention to best practice.

One might think that methodological consciousness-raising is unnecessary, because on any particular issue good arguments will tend to drive out bad in the long run, by a reverse analogue of Gresham’s Law. But that is over-optimistic. Very often – not least in debates between realists and anti-realists – a philosopher profoundly *wants* one answer rather than another to a philosophical question to be right, and is therefore predisposed to accept arguments that go in the preferred direction and reject contrary ones. Where the level of obscurity is high, wishful thinking may be more powerful than the ability to distinguish good arguments from bad, to the point that convergence in the evaluation of arguments never occurs.

Consider a dispute between rival theories in natural science. Each theory has its committed defenders, who have invested much time, energy, and emotion in its survival. The theories are not empirically equivalent, but making an empirical determination between them requires experimental skills of a high order. We may predict that if the standards of accuracy and conscientiousness in the community are high enough, truth will eventually triumph. But if the community is slightly more tolerant of sloppiness and rhetorical obfuscation, then

each school may be able to survive indefinitely, claiming empirical vindication and still verbally acknowledging the value of rigor, by protecting samples from impurities a little less adequately, describing experimental results a little more tendentiously, giving a little more credit to *ad hoc* hypotheses, dismissing opposing arguments as question-begging a little more quickly, and so on. Each tradition maintains recruitment by its dominance and prestige in some departments or regions. A small difference in how carefully standards are applied can make the large difference between eventual convergence and ultimate divergence.

It seems likely that some parts of contemporary analytic philosophy just pass the methodological threshold for some cumulative progress to occur, however slowly, while others fall short of the threshold. For example, a reasonable fear is that debates over realism and anti-realism fall short. That is not to condemn every piece of work in such areas individually – which would surely be unfair – but to say that collectively the community of participants has not held itself responsible to high enough methodological standards. Perhaps these debates raise even more difficult issues than are encountered elsewhere in philosophy: if so, all the more reason to apply the very highest standards available. As already noted, that appears not to have happened.

How can we do better? We can make a useful start by getting the simple things right. Much even of analytic philosophy moves too fast in its haste to reach the sexy bits. Details are not given the care they deserve: crucial claims are vaguely stated, significantly different formulations are treated as though they were equivalent, examples are under-described, arguments are gestured at rather than properly made, their form is left unexplained, and so on. A few resultant errors easily multiply to send inquiry in completely the wrong direction. Shoddy work is sometimes masked by pretentiousness, allusiveness, gnomic concision, or winning informality. But often there is no special disguise: producers and consumers have simply not taken enough trouble to check the details. We need the unglamorous virtue of patience to read and write philosophy that is as perspicuously structured as the difficulty of the subject requires, and the austerity to be dissatisfied with appealing prose that does not meet those standards. The fear of boring oneself or one's readers is a great enemy of truth. Pedantry is a fault on the right side.

Precision is often regarded as a hyper-cautious characteristic. It is importantly the opposite. Vague statements are the hardest to convict of error. Obscurity is the oracle's self-defense. To be precise is to make it as easy as possible for others to prove one wrong. That is what requires courage. But the community can lower the cost of precision by keeping in mind that precise errors often do more than vague truths for scientific progress.

Would it be a good bargain to sacrifice depth for rigor? That bargain is not on offer in philosophy, any more than it is in mathematics. No doubt, if we aim to be rigorous, we cannot expect to sound like Heraclitus, or even Kant: we have to sacrifice the stereotype of depth. Still, it is rigor, not its absence, that prevents one from sliding over the deepest difficulties, in an agonized rhetoric of profundity. Rigor and depth both matter: but while the continual deliberate pursuit of rigor is a good way of achieving it, the continual deliberate pursuit of depth (as of happiness) is far more likely to be self-defeating. Better to concentrate on trying to say something true and leave depth to look after itself.

Nor are rigor and precision enemies of the imagination, any more than they are in mathematics. Rather, they increase the demands on the imagination, not least by forcing one to imagine examples with exactly the right structure to challenge a generalization; cloudiness will not suffice. They make imagination consequential in a way in which it is not in their absence. The most rigorous and precise discussion often involves the most playfulness and laughter: toying with subtly different combinations of ideas yields surprising scenarios. Humorless solemnity masks sloppiness and confusion.

Beyond rigor and precision, mathematics has less obvious values to teach. In particular, a mathematical training makes one appreciate the importance of the aesthetics of definitions. Experience shows that a mathematician or logician with no ability to discriminate between fruitful and unfruitful definitions is unlikely to achieve much in research. Such discriminations involve a sort of aesthetic judgment. The ugly, convoluted, ramshackle definitions of concepts and theses that philosophers seem to feel no shame in producing are of just the kind to strike a mathematician as pointless and sterile. Of course, it is notoriously hard to explain *why* aesthetic criteria are a good methodological guide, but it would be dangerously naïve to abandon them for that reason.

In addition to the humdrum methodological virtues, we need far more reflectiveness about how philosophical debates are to be subjected to enough constraints to be worth conducting. For example, Dummettian anti-realism about the past involved, remarkably, the abandonment of two of the main constraints on much philosophical activity. In rejecting instances of the law of excluded middle concerning past times, such as “Either a mammoth stood on this spot a hundred thousand years ago or no mammoth stood on this spot a hundred thousand years ago,” the anti-realist rejected both common sense and classical logic. Those constraints are simultaneously abandoned in many contemporary philosophical debates too, for example over vagueness. Neither constraint is methodologically sacrosanct; both can intelligibly be challenged, even together. But when participants in a debate are allowed to throw out both simultaneously, methodological alarm bells should ring: it is at least not obvious that enough constraints are left to frame a fruitful discussion. Yet such qualms surface remarkably little (although Dummett himself did not ignore the methodological issues).

Part of the problem is that it is often left unclear just how extensively a constraint is being challenged. A philosopher treats the law of excluded middle as if it carried no authority whatsoever but implicitly relies on other logical principles (perhaps in the metalanguage): exactly which principles of logic are supposed to carry authority? A philosopher treats some common sense judgment as if it carried no authority whatsoever but implicitly relies on other judgments that are found pre-philosophically obvious: exactly which such judgments are supposed to carry authority?

When law and order break down, the result is not freedom or anarchy but the capricious tyranny of petty feuding warlords. Similarly, the unclarity of constraints in philosophy leads to authoritarianism. Whether an argument is widely accepted depends not on publicly accessible criteria that we can all apply for ourselves but on the say-so of charismatic authority figures. Pupils cannot become autonomous from their teachers because they cannot securely learn the standards by which their teachers judge. A modicum of willful unpredictability in the application of standards is a good policy for a professor who does not want his students to gain too much independence. Although intellectual deference is not always a bad thing,

some debates have seen far too much of it. We can reduce it by articulating and clarifying the constraints.

Philosophy can never be reduced to mathematics. But we can often produce mathematical models of fragments of philosophy and, when we can, we should. No doubt the models usually involve wild idealizations. It is still progress if we can agree what consequences an idea has in one very simple case. Many ideas in philosophy do not withstand even that very elementary scrutiny, because the attempt to construct a non-trivial model reveals a hidden structural incoherence in the idea itself. By the same token, an idea that does not collapse in a toy model has at least something going for it. Once we have an unrealistic model, we can start worrying how to construct less unrealistic models.

Philosophers who reject the constraints mentioned above can say what constraints they would regard as appropriate. Of course, those who deny that philosophy is a theoretical discipline at all may reject the very idea of such constraints. But surely the best way to test the theoretical ambitions of philosophy is to go ahead and try to realize them in as disciplined a way as possible. If the anti-theorists can argue convincingly that the long-run results do not constitute progress, that is a far stronger case than is a preconceived argument that no such activity could constitute progress. On the other hand, if they cannot argue convincingly that the long-run results do not constitute progress, how is their opposition to philosophical theory any better than obscurantism?

Unless names are invidiously named, sermons like this one tend to cause less offence than they should, because everyone imagines that they are aimed at other people. Those who applaud a methodological platitude usually assume that they comply with it. I intend no such comfortable reading. To one degree or another, we all fall short not just of the ideal but of the desirable and quite easily possible. Certainly this afterword exhibits hardly any of the virtues that it recommends, although with luck it may still help a bit to propagate those virtues (do as I say, not as I do). Philosophy has never been done for an extended period according to standards as high as those that are now already available, if only the profession will take them seriously to heart. None of us knows how far we can get by applying them systematically enough for long enough. We can find out only by trying.

In making these comments, it is hard not to feel like the headmaster of a minor public school at speech day, telling everyone to pull their socks up after a particularly bad term. It is therefore appropriate to end with a misquotation from Winston Churchill. This is not the end of philosophy. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.

# Appendix 1

## Modal Logic within Counterfactual Logic

---

This appendix sketches the development of logics of possibility and necessity as subsystems of logics of the counterfactual conditional, on suitable definitions of the former in terms of the latter. No particular formal semantic account of the counterfactual conditional is assumed, although various sorts of model theory are occasionally used in auxiliary roles. The emphasis is on questions of deducibility from principles plausible on an informal reading of the counterfactual conditional.

For most purposes our object language is  $L$ , which has countably many propositional variables  $p, q, r, \dots$ , the propositional constant  $\perp$  (a logical falsehood) and two binary connectives,  $\rightarrow$  (the material conditional) and  $\Box\rightarrow$  (the counterfactual conditional). Other truth-functional operators are introduced as metalinguistic abbreviations in the usual way; for example,  $\neg A$  is  $A \rightarrow \perp$ . The metalinguistic variables “ $A$ ,” “ $B$ ,” “ $C$ ,”  $\dots$  range over all formulas.

Except when otherwise specified, we work in the following axiomatic system ( $\vdash$  means theoremhood):

PC	If $A$ is a truth-functional tautology then $\vdash A$
REFLEXIVITY	$\vdash A \Box\rightarrow A$
VACUITY	$\vdash (\neg A \Box\rightarrow A) \rightarrow (B \Box\rightarrow A)$
MP	If $\vdash A \rightarrow B$ and $\vdash A$ then $\vdash B$
CLOSURE	If $\vdash (B_1 \& \dots \& B_n) \rightarrow C$ then $\vdash ((A \Box\rightarrow B_1) \& \dots \& (A \Box\rightarrow B_n)) \rightarrow (A \Box\rightarrow C)$
EQUIVALENCE	If $\vdash A \equiv A^*$ then $\vdash (A \Box\rightarrow B) \equiv (A^* \Box\rightarrow B)$

These axiom schemas and inference rules constitute a weak subsystem of David Lewis’s “official” logic of counterfactuals, VC (1986:

132). PC, REFLEXIVITY, and VACUITY are his axiom schemas (1), (3), and (4) respectively, and MP is his rule of Modus Ponens (for  $\rightarrow$ ). CLOSURE is his rule of Deduction within Conditionals (unlike Lewis, we allow  $n = 0$ , interpreting this case as the rule that if  $\vdash C$  then  $\vdash A \square\rightarrow C$ ; but that special case is anyway derivable from CLOSURE for  $n = 1$  and REFLEXIVITY). EQUIVALENCE is a special case of Lewis's rule of Interchange of Logical Equivalents (incorrectly omitted from the original 1973 edition (1986: ix)); Interchange of Logical Equivalents in its full generality for all sentential contexts in L is derivable from EQUIVALENCE, CLOSURE, PC, and MP (proof: by induction on the construction of formulas).

PC and MP simply encapsulate the background classical logic. REFLEXIVITY reflects the triviality that in developing a counterfactual supposition we can start with that supposition itself. The point of VACUITY is that  $\neg A$  is the “worst” antecedent for A as consequent; if A is forthcoming even in that case, it is forthcoming in every case. To think of it another way,  $\neg A \square\rightarrow A$  can be true only by being vacuously true, in which case A is true in every eventuality. CLOSURE means that in developing a counterfactual supposition, we can include any logical consequence of the results obtained so far. EQUIVALENCE goes with the idea that differences between logically equivalent counterfactual suppositions are in effect differences only in the mode of presentation of the way things are being supposed to be.

One way in which the present subsystem of Lewis's system is weak is that it lacks his irredundant “centering” axiom schema (7)  $(A \& B) \rightarrow (A \square\rightarrow B)$ , for, unlike the principles above, it is invalid when  $\square\rightarrow$  is reinterpreted as strict implication in S5. It also lacks his “weak centering” axiom schema (6)  $(A \square\rightarrow B) \rightarrow (A \rightarrow B)$ , the addition of which will be considered later. Finally, our subsystem lacks the axiom schema (5) for whose length and obscurity Lewis apologizes:

$$(A \square\rightarrow \neg B) \vee (((A \& B) \square\rightarrow C) \equiv (A \square\rightarrow (B \rightarrow C)))$$

(Lewis 1986: 133). Unlike (6) and (7), (5) is part of Lewis's core system V. We can check that (5) is irredundant in Lewis's axiomatization by considering an unintended semantics on which it is invalid but all his other axiom schemas are valid and his rules preserve validity. Specifically, suppose that each model supplies a set of worlds and a func-

tion  $f$  from formulas  $A$  and worlds  $w$  to sets of worlds  $f(A, w)$  satisfying the constraints (i)  $A$  is true at every world in  $f(A, w)$ ; (ii)  $f(A, w)$  is empty only if  $A$  is true at no world; (iii) if  $A$  is true at  $w$  then  $f(A, w) = \{w\}$ ; (iv) if  $A$  and  $B$  are true at exactly the same worlds then  $f(A, w) = f(B, w)$ . The semantic clause for  $\Box \rightarrow$  is then that  $A \Box \rightarrow B$  is true at  $w$  if and only  $B$  is true at every world in  $f(A, w)$ . One easily checks by induction on the length of proofs that all Lewis's other axiom schemas are true at every world in every model under this semantics, and that his rules preserve this property. However, schema (5) fails. Consider a set of four worlds  $\{0, 1, 2, 3\}$ , let the atomic formula  $p$  be true at 1, 2 and 3 only,  $q$  be true at 1 and 2 only,  $r$  be true at 1 only; if  $A$  is true at  $w$  let  $f(A, w)$  be  $\{w\}$ ; if  $A$  is false at  $w$  let  $f(A, w)$  be the set of all worlds at which  $A$  is true, except when  $w$  is 0 and  $A$  is true at 1 and 2 only, in which case  $f(A, w)$  is  $\{1\}$ . These stipulations obviously satisfy (i)–(iv). Now  $p \Box \rightarrow \neg q$  is false at 0, because  $f(p, 0)$  is  $\{1, 2, 3\}$  and  $\neg q$  is false at 1 and 2, and  $p \Box \rightarrow (q \rightarrow r)$  is false at 0 because  $q \rightarrow r$  is false at 2, but  $(p \& q) \Box \rightarrow r$  is true at 0, because  $f(p \& q, 0)$  is  $\{1\}$  and  $r$  is true at 1. Consequently,  $(p \Box \rightarrow \neg q) \vee (((p \& q) \Box \rightarrow r) \equiv (p \Box \rightarrow (q \rightarrow r)))$  is false at 0. In this setting, we therefore cannot apply most of Lewis's results about derived modal logics within counterfactual logics (1986: 137–42), because they depend on completeness theorems for his counterfactual logics with respect to classes of models with respect to which the present systems are incomplete. Not that any reason has been provided to regard Lewis's extra schemas as informally invalid on their intended natural language readings (if we do not already assume the correctness of Lewis's semantic theory); the point is just that their informal validity on those readings is hard to assess, so it is better to derive modal logic from counterfactual logic without them.

CLOSURE and EQUIVALENCE are not quite as straightforward as they look. In a language with a rigidifying “actually” operator  $@$ ,  $p \equiv @p$  is arguably a logical truth. But if it is a theorem, each of CLOSURE and EQUIVALENCE separately (when combined with REFLEXIVITY) yields the theorem  $p \Box \rightarrow @p$ , which is false on many interpretations: “If it had rained, it would have actually rained” is false if it did not rain. In the terminology of Davies and Humberstone (1980), CLOSURE and EQUIVALENCE preserve general validity (truth at every world of every model) but not real world validity (truth at the actual world of every model). Thus CLOSURE and EQUIVALENCE must be restricted to theorems derived solely by

appeal to axioms and rules that preserve general validity. A similar restriction is needed on the standard Rule of Necessitation (RN) in modal logic, that if  $A$  is a theorem so is  $\Box A$ , for even if  $p \rightarrow @p$  is logically true,  $\Box(p \rightarrow @p)$  may be false. For present purposes we can ignore this complication, since the languages under consideration lack operators such as “actually” (see Williamson (2006a) for further discussion).

For our immediate purposes, we expand  $L$  to the language  $L^+$  by adding propositional quantifiers. That is, if  $p$  is a propositional variable and  $A$  is a formula of  $L^+$ ,  $\forall p A$  is also a formula of  $L^+$ . We extend the axiomatization by a corresponding axiom schema and rule (where  $A[B/p]$  is the result of substituting the formula  $B$  for all free occurrences of  $p$  in  $A$ , on the assumption that no variable free in  $B$  thereby becomes bound):

UINST If  $p$  is any propositional variable then  $\vdash \forall p A \rightarrow A[B/p]$   
UGEN If  $p$  is any propositional variable not free in  $A$ , and  
 $\vdash A \rightarrow B$  then  $\vdash A \rightarrow \forall p B$

This system, like that for  $L$ , satisfies the rule of substitution of proved material equivalents, in the sense that if  $\vdash B \equiv B^*$  then  $\vdash A[B/p] \equiv A[B^*/p]$  for any formula  $A$  and propositional variable  $p$  (proof: by induction on the complexity of  $A$ ). Thus proved material equivalents are interchangeable in all relevant contexts. In the setting of possible worlds semantics, UINST and UGEN are sound when the propositional quantifiers are interpreted as ranging over all subsets of the set of possible worlds associated with the given model, but they will not yield a complete system, since they do not guarantee the existence of maximally specific possible propositions, true in exactly one world (for example, one cannot derive  $\exists p (p \ \& \ \forall q (q \rightarrow \Box(p \rightarrow q)))$ ).<sup>1</sup> For present purposes, those stronger assumptions are unnecessary.

<sup>1</sup> See the pioneering works of Fine (1970) and Kaplan (1970) for more technical detail on propositional quantification in modal logic. Williamson (1999a) discusses its interpretation: interpreting it by means of quantification into name position in the metalanguage, over sets of possible worlds or anything else, is arguably only a rough approximation to its philosophically most significant interpretation, which involves ineliminable quantification into sentence position.

Our first task is to show that three candidate definitions of  $\Box A$  in  $L^+$  are mutually equivalent: (i)  $\forall p (p \Box \rightarrow A)$  (where  $p$  is not free in  $A$ ); (ii)  $\neg A \Box \rightarrow A$ ; (iii)  $\neg A \Box \rightarrow \perp$ . First we establish the equivalence of (i) and (ii):

(1) $\forall p (p \Box \rightarrow A) \rightarrow (\neg A \Box \rightarrow A)$	UINST
(2) $(\neg A \Box \rightarrow A) \rightarrow (p \Box \rightarrow A)$	VACUITY
(3) $(\neg A \Box \rightarrow A) \rightarrow \forall p (p \Box \rightarrow A)$	2, UGEN
(4) $\forall p (p \Box \rightarrow A) \equiv (\neg A \Box \rightarrow A)$	1, 3, PC, MP

Now we establish the equivalence of (ii) and (iii):

(1) $(A \& \neg A) \rightarrow \perp$	PC
(2) $((\neg A \Box \rightarrow A) \& (\neg A \Box \rightarrow \neg A)) \rightarrow (\neg A \Box \rightarrow \perp)$	1, CLOSURE
(3) $\neg A \Box \rightarrow \neg A$	REFLEXIVITY
(4) $(\neg A \Box \rightarrow A) \rightarrow (\neg A \Box \rightarrow \perp)$	2, 3, PC, MP
(5) $\perp \rightarrow A$	PC
(6) $(\neg A \Box \rightarrow \perp) \rightarrow (\neg A \Box \rightarrow A)$	5, CLOSURE
(7) $(\neg A \Box \rightarrow A) \equiv (\neg A \Box \rightarrow \perp)$	4, 6, PC, MP

Thus (i), (ii), and (iii) are mutually interchangeable in all relevant contexts. It matters little which of them we use to define  $\Box A$ . However, the complexities of propositional quantification are best avoided when not needed, and (iii) is marginally simpler than (ii), so we treat  $\Box A$  as a metalinguistic abbreviation for  $\neg A \Box \rightarrow \perp$ . We therefore return to the propositional language  $L$ , and omit the quantifier rules. As usual in modal logic we treat  $\Diamond A$  as a metalinguistic abbreviation for  $\neg \Box \neg A$ , which in our case is  $\neg(\neg \neg A \Box \rightarrow \perp)$ , which is equivalent by EQUIVALENCE to  $\neg(A \Box \rightarrow \perp)$ .

The next task is to check the status on our definitions of two principles used in the main text:

NECESSITY  $\vdash \Box(A \rightarrow B) \rightarrow (A \Box \rightarrow B)$   
 POSSIBILITY  $\vdash (A \Box \rightarrow B) \rightarrow (\Diamond A \rightarrow \Diamond B)$

We prove them in our system as follows. First, NECESSITY:

(1) $\Box(A \rightarrow B) \rightarrow (\neg(A \rightarrow B) \Box \rightarrow \perp)$	DEF $\Box$ , PC
(2) $(\neg(A \rightarrow B) \Box \rightarrow \perp) \rightarrow (\neg(A \rightarrow B) \Box \rightarrow (A \rightarrow B))$	PC, CLOSURE

(3)	$(\neg(A \rightarrow B) \square \rightarrow (A \rightarrow B)) \rightarrow (A \square \rightarrow (A \rightarrow B))$	VACUITY
(4)	$\square(A \rightarrow B) \rightarrow (A \square \rightarrow (A \rightarrow B))$	1, 2, 3, PC, MP
(5)	$((A \square \rightarrow (A \rightarrow B)) \& (A \square \rightarrow A)) \rightarrow (A \square \rightarrow B)$	PC, CLOSURE
(6)	$(A \square \rightarrow (A \rightarrow B)) \rightarrow (A \square \rightarrow B)$	5, REFLEXIVITY, PC, MP
(7)	$\square(A \rightarrow B) \rightarrow (A \square \rightarrow B)$	4, 6, PC, MP

Then, POSSIBILITY:

(1)	$\neg\Diamond B \rightarrow (\neg\neg B \square \rightarrow \perp)$	DEF $\Diamond$ , PC
(2)	$(\neg\neg B \square \rightarrow \perp) \rightarrow (\neg\neg B \square \rightarrow \neg B)$	PC, CLOSURE
(3)	$(\neg\neg B \square \rightarrow \neg B) \rightarrow (A \square \rightarrow \neg B)$	VACUITY
(4)	$\neg\Diamond B \rightarrow (A \square \rightarrow \neg B)$	1, 2, 3, PC, MP
(5)	$((A \square \rightarrow B) \& (A \square \rightarrow \neg B)) \rightarrow (A \square \rightarrow \perp)$	PC, CLOSURE
(6)	$((A \square \rightarrow B) \& \neg\Diamond B) \rightarrow (A \square \rightarrow \perp)$	4, 5, PC, MP
(7)	$(\neg\neg A \square \rightarrow \perp) \rightarrow \neg\Diamond A$	DEF $\Diamond$ , PC
(8)	$(A \square \rightarrow \perp) \supset (\neg\neg A \square \rightarrow \perp)$	PC, EQUIVALENCE
(9)	$((A \square \rightarrow B) \& \neg\Diamond B) \rightarrow \neg\Diamond A$	6, 7, 8, PC, MP
(10)	$(A \square \rightarrow B) \rightarrow (\Diamond A \rightarrow \Diamond B)$	9, PC, MP

We now turn to deriving some basic principles of modal logic within counterfactual logic. The weakest normal modal logic is K, which is axiomatized by PC, MP, and the following axiom schema and rule:

K  $\vdash \square(A \rightarrow B) \rightarrow (\square A \rightarrow \square B)$

RN If  $\vdash A$  then  $\vdash \square A$

We derive K in our system thus:

(1)	$\square A \rightarrow (\neg A \square \rightarrow \perp)$	PC, DEF $\square$
(2)	$\square A \rightarrow (\neg A \square \rightarrow A)$	1, PC, CLOSURE, MP
(3)	$\square A \rightarrow (\neg B \square \rightarrow A)$	2, VACUITY, PC, MP
(4)	$\square(A \rightarrow B) \rightarrow (\neg B \square \rightarrow (A \rightarrow B))$	Like 3
(5)	$((\neg B \square \rightarrow (A \rightarrow B)) \& (\neg B \square \rightarrow A)) \rightarrow (\neg B \square \rightarrow B)$	PC, CLOSURE
(6)	$(\square(A \rightarrow B) \& \square A) \rightarrow (\neg B \square \rightarrow B)$	3, 4, 5, PC, MP

---

(7) $(\neg B \rightarrow B) \rightarrow (\neg B \rightarrow \perp)$	REFLEXIVITY, CLOSURE, PC, MP
(8) $(\neg B \rightarrow B) \rightarrow \Box B$	7, DEF $\Box$
(9) $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$	6, 8, PC, MP

Here is a derivation of RN:

(1) $A$	Theorem by assumption
(2) $\neg A \rightarrow \perp$	1, PC, MP
(3) $(\neg A \rightarrow \neg A) \rightarrow (\neg A \rightarrow \perp)$	2, CLOSURE
(4) $\neg A \rightarrow \perp$	3, REFLEXIVITY, PC, MP
(5) $\Box A$	4, DEF $\Box$

Thus all theorems of K are theorems of our system, under our definition of  $\Box$ .

We can prove something stronger: the modal principles derivable in our current system are *just* those derivable in K. More precisely, let  $L_{\Box}$  be the language of propositional modal logic, built up from the propositional variables,  $\perp$ ,  $\rightarrow$  and  $\Box$  (treated as primitive). Let  $*$  be the mapping from  $L_{\Box}$  to L that corresponds to our definition of  $\Box$ :

- $*p = p$  for each propositional variable  $p$
- $*\perp = \perp$
- $*(A \rightarrow B) = *A \rightarrow *B$
- $*\Box A = \neg *A \rightarrow \perp$

Then for any formula  $A$  of  $L_{\Box}$ ,  $*A$  is a theorem of our system ( $\vdash *A$ ) if and only if  $A$  is a theorem of K ( $\vdash_K A$ ). We have in effect already proved that if  $\vdash_K A$  then  $\vdash *A$ . The converse is trickier, because the proof of  $*A$  in our system may involve formulas such as  $p \rightarrow q$  that are not of the form  $*B$  for any formula  $B$  of  $L_{\Box}$ . We define an auxiliary “unintended” mapping  $\hat{\cdot}$  back from L to  $L_{\Box}$ :

- $\hat{p} = p$  for each propositional variable  $p$
- $\hat{\perp} = \perp$
- $\hat{(A \rightarrow B)} = \hat{A} \rightarrow \hat{B}$
- $\hat{(\Box A)} = \Box(\hat{A} \rightarrow \hat{B})$

We note two easy lemmas.

- (I) For any formula  $A$  of  $L$ , if  $\vdash A$  then  $\vdash_k {}^A$ . Proof: by induction on the length of proofs in our system.
- (II) For any formula  $A$  of  $L_{\square}$ ,  $\vdash_k A \equiv {}^*A$ . Proof: by induction on the complexity of  $A$ .

Now suppose that  $A$  is a formula of  $L_{\square}$  and  $\vdash {}^*A$ . By (I),  $\vdash_k {}^*A$ . By (II),  $\vdash_k A \equiv {}^*A$ . Therefore  $\vdash_k A$ , as required. Thus  $\vdash {}^*A$  if and only if  $\vdash_k A$ .

The system K is far too weak to be an adequate logic of metaphysical possibility and necessity. The most saliently missing principle is that what is necessarily so is so:

$$T \vdash \square A \rightarrow A$$

We can derive T in our system by adding Lewis's "weak centering" principle (schema (6) in his official logic of counterfactuals (1986: 132); it is also axiom schema (a6) in Stalnaker 1968), which corresponds to modus ponens for the counterfactual conditional given the logic of the material conditional:

$$MP\square\rightarrow \vdash (A \square\rightarrow B) \rightarrow (A \rightarrow B)$$

T is an immediate consequence of  $MP\square\rightarrow$ :

- (1)  $(\neg A \square\rightarrow \perp) \rightarrow (\neg A \rightarrow \perp)$   $MP\square\rightarrow$
- (2)  $(\neg A \square\rightarrow \perp) \rightarrow A$  1, PC, MP
- (3)  $\square A \rightarrow A$  2, DEF $\square$

By a proof along just the same lines as for K (with the same mappings), we can show that for any formula  $A$  of  $L_{\square}$ ,  ${}^*A$  is a theorem of our system extended by  $MP\square\rightarrow$  if and only if  $A$  is a theorem of KT, the result of extending K (as axiomatized above) by T. Thus PC, REFLEXIVITY, VACUITY,  $MP\square\rightarrow$ , MP, CLOSURE, and EQUIVALENCE induce the simple logic KT for metaphysical modality.

$MP\square\rightarrow$  is an immensely plausible principle. If we discover that  $e$  happened without  $f$ , doesn't that refute the claim that if  $e$  had hap-

pened,  $f$  would have happened?<sup>2</sup> Nevertheless, it is worth observing that the full strength of  $MP\Box\rightarrow$  is not needed to derive  $T$ . For if we merely add  $T$  itself to our original system (read by means of  $DEF\Box$ ), we cannot derive  $MP\Box\rightarrow$ . We can show this by giving an unintended model theory that validates PC, REFLEXIVITY, VACUITY, T, MP, CLOSURE, and EQUIVALENCE but not  $MP\Box\rightarrow$ . It is a “possible worlds” semantics, but with the natural numbers playing the role of the worlds. The clause for  $\Box\rightarrow$  is this:  $A \Box\rightarrow B$  is true at all worlds iff either  $A$  is false at all worlds or  $B$  is true at the least world at which  $A$  is true (“least” in the sense of the usual ordering of the natural numbers; recall that every nonempty set of natural numbers has a least member); otherwise  $A \Box\rightarrow B$  is false at all worlds. Everything else is standard. It is routine to check (by induction on the length of proofs) that every formula of  $L$  derivable from PC, REFLEXIVITY, VACUITY, T, MP, CLOSURE, and EQUIVALENCE is true at all worlds in all such models. For example, in the case of  $T$ , suppose that  $\Box A$  is true at a world, which is to say that  $\neg A \Box\rightarrow \perp$  is true at that world; since  $\neg A \Box\rightarrow \perp$  cannot be non-vacuously true, it must be vacuously true; thus  $A$  is true at every world. But not all instances of  $MP\Box\rightarrow$  are true at all worlds in all such models. For example, let  $p$  be true at 0 but false at every other world. Then  $\neg \perp \rightarrow p$  is true at every world, while  $\neg \perp \rightarrow p$  is false at 1.

A more controversial but still plausible principle about metaphysical modality is the characteristic axiom schema of the modal system S5, known as E:

$$E \vdash \Diamond A \rightarrow \Box \Diamond A$$

KTE is simply S5; in that system, matters of possibility and necessity are always non-contingent. We can also derive in it the characteristic principle of S4:

<sup>2</sup> One can accept a counterfactual when rationally unwilling to apply modus ponens to it, in the sense that on learning its antecedent one would reject the counterfactual rather than accept its consequent. For example, I accept “If Oswald had not shot Kennedy, Kennedy would not have been shot,” but if I come to accept “Oswald did not shoot Kennedy,” I will not conclude “Kennedy was not shot.” But that is no threat to the validity of modus ponens. In circumstances in which both “If Oswald had not shot Kennedy, Kennedy would not have been shot” and “Oswald did not shoot Kennedy” are true, so is “Kennedy was not shot.”

4S  $\vdash \Box A \rightarrow \Box\Box A$

(Hughes and Cresswell (1996) provides appropriate background in modal logic.) If we read E directly in terms of our counterfactual definitions of the modal operators,  $\Box\Diamond A$  becomes a counterfactual conditional with a (negated) counterfactual conditional in its antecedent, which is quite hard to get a feel for. Lewis adds axioms involving such intractable counterfactuals to his system to obtain S5. Here is a more natural equivalent of E in counterfactual conditional terms:

ES  $\vdash (A \Box\rightarrow (B \Box\rightarrow \perp)) \rightarrow ((A \Box\rightarrow \perp) \vee (B \Box\rightarrow \perp))$

The embedded counterfactual conditional has been moved into the consequent, where such embeddings occur somewhat more naturally. Informally, ES says that embedding one possible counterfactual hypothesis inside another cannot lead to an impossibility: even if B is incompatible with A, counterfactually supposing B within the counterfactual supposition of A takes one back out of the A worlds into the B worlds, not to an impossibility.

The generalization of ES to arbitrary sentences in place of the logical falsehood is much less plausible:

ES+  $\vdash (A \Box\rightarrow (B \Box\rightarrow C)) \rightarrow ((A \Box\rightarrow C) \vee (B \Box\rightarrow C))$

If I had been a French grocer then I would have been such that if I had been a philosopher I would have been a French philosopher; but it is not the case that if I had been a French grocer I would have been a French philosopher, nor is it the case that if I had been a philosopher I would have been a French philosopher. In terms of Lewis's similarity semantics, suppose that  $p$  holds only at the counterfactual world  $w$ ,  $q$  holds only at the actual world and a third world  $x$ , closer to  $w$  than the actual world is, and  $r$  holds only at  $x$ . Then  $w$  is a  $q \Box\rightarrow r$  world, because the closest  $q$  world to  $w$  is  $x$ , which is an  $r$  world; thus the actual world is a  $p \Box\rightarrow (q \Box\rightarrow r)$  world, since  $w$  is the closest  $p$  world to the actual world; but the actual world is neither a  $p \Box\rightarrow r$  world (since the closest  $p$  world to the actual world,  $w$ , is not an  $r$  world) nor a  $q \Box\rightarrow r$  world (since the closest  $q$  world to the actual world is the actual world itself, which is not an  $r$  world). Thus

ES+ is invalid in Lewis's semantics. By contrast, ES holds on Lewis's semantics provided that all worlds form a single similarity space (compare Lewis's uniformity condition (1986: 120–1)). For then  $B \Box \rightarrow \perp$  is false at every world if  $B$  is true at some world; thus if  $B \Box \rightarrow \perp$  is false at a world, so  $B$  is true at some world, then  $B \Box \rightarrow \perp$  is true at exactly the same worlds as  $\perp$ , so  $A \Box \rightarrow (B \Box \rightarrow \perp)$  and  $A \Box \rightarrow \perp$  have the same truth-value at all worlds; thus ES holds (consequently, ES does not entail ES+). The plausibility of ES depends on the occurrence of a logical falsehood in the consequent. Although we will not attempt to determine here whether ES should ultimately be accepted, it at least gives us a new perspective on the status of S5 (Salmon 1982: 238–40, 1989, and 1993 argue that S4 and therefore S5 are invalid for metaphysical modality; Williamson 1990: 126–43 and 2000a: 119–20 reply).

We still have to establish the equivalence of ES with E. First, we argue from ES to E in our original system:

- (1)  $((\neg\neg A \Box \rightarrow \perp) \Box \rightarrow (\neg\neg A \Box \rightarrow \perp)) \rightarrow$  ES
- $((\neg\neg A \Box \rightarrow \perp) \Box \rightarrow \perp) \vee (\neg\neg A \Box \rightarrow \perp)$
- (2)  $((\neg\neg A \Box \rightarrow \perp) \Box \rightarrow \perp) \vee (\neg\neg A \Box \rightarrow \perp)$  1, REFLEXIVITY, MP
- (3)  $\neg(\neg\neg A \Box \rightarrow \perp) \rightarrow$  2, EQUIVALENCE,  
 $(\neg\neg (\neg\neg A \Box \rightarrow \perp) \Box \rightarrow \perp)$  PC, MP
- (4)  $\Diamond A \rightarrow \Box\Diamond A$  3, DEF $\Diamond$ , DEF $\Box$

Now we establish the converse, again in our original system:

- (1)  $\Diamond B \rightarrow \Box\Diamond B$  E
- (2)  $\neg(B \Box \rightarrow \perp) \rightarrow (\neg\neg(B \Box \rightarrow \perp) \Box \rightarrow \perp)$  1, EQUIVALENCE,  
PC, MP, DEF $\Diamond$ , DEF $\Box$
- (3)  $(\neg\neg(B \Box \rightarrow \perp) \Box \rightarrow \perp) \rightarrow$  CLOSURE, MP, PC  
 $(\neg\neg(B \Box \rightarrow \perp) \Box \rightarrow \neg(B \Box \rightarrow \perp))$
- (4)  $\neg(B \Box \rightarrow \perp) \rightarrow (\neg\neg(B \Box \rightarrow \perp) \Box \rightarrow \neg(B \Box \rightarrow \perp))$  2, 3, MP, PC
- (5)  $\neg(B \Box \rightarrow \perp) \rightarrow (A \Box \rightarrow \neg(B \Box \rightarrow \perp))$  4, VACUITY, MP, PC
- (6)  $((A \Box \rightarrow (B \Box \rightarrow \perp)) \& (A \Box \rightarrow \neg(B \Box \rightarrow \perp)) \rightarrow$   
 $(A \Box \rightarrow \perp)$  CLOSURE, MP, PC
- (7)  $(A \Box \rightarrow (B \Box \rightarrow \perp)) \rightarrow$  5, 6, MP, PC  
 $((A \Box \rightarrow \perp) \vee (B \Box \rightarrow \perp))$

Although PC, REFLEXIVITY, VACUITY,  $\text{MP}\Box\rightarrow$ , ES, MP, CLOSURE, and EQUIVALENCE together yield the full strength of S5, they still constitute a rather weak logic of counterfactuals. For example, they do not yield axiom schema (a7) from Stalnaker (1968), a strengthening of EQUIVALENCE:

$$(a7) \vdash ((A \Box\rightarrow B) \& (B \Box\rightarrow A)) \rightarrow ((A \Box\rightarrow C) \rightarrow (B \Box\rightarrow C))$$

To check independence, consider another deviant semantics in which the possible worlds are the natural numbers. Let  $A \Box\rightarrow B$  be true at a world  $w$  if and only if three conditions hold: (i) if  $A$  is true at  $w$  then  $B$  is true at  $w$ ; (ii) if  $A$  is true at exactly one world then  $B$  is also true at that world; (iii) if  $A$  is true at a world  $x$  and at some world  $y$  such that  $x > y$  then  $B$  is true at  $x$ . In particular, therefore,  $A \Box\rightarrow \perp$  is true at all worlds if  $A$  is false at all worlds; otherwise  $A \Box\rightarrow \perp$  is false at all worlds. Everything else is standard. It is routine to check that all theorems of our system are true in all such models. But (a7) fails: for if  $p$  is true at just 1 and 2,  $q$  at just 0, 1 and 2 and  $r$  just at 2, then  $p \Box\rightarrow q$ ,  $q \Box\rightarrow p$  and  $p \Box\rightarrow r$  are true but  $q \Box\rightarrow r$  false at 2 (since  $q$  is true and  $r$  false at 1, which is not the least world at which  $q$  is true). The same semantics shows that the complex axiom schema (5) of Lewis's official system VC (1986: 132) is not derivable in our system (since  $(q \& p) \Box\rightarrow r$  is true but  $q \Box\rightarrow (p \rightarrow r)$  is false at 2). We might wish to add some of these further principles to our system.

Moderately natural counterfactual equivalents of other modal principles can also be provided. For example, the 4 schema  $\vdash \Box A \rightarrow \Box\Box A$  is equivalent to this schema:

$$4S \vdash (A \Box\rightarrow \perp) \rightarrow (A \Box\rightarrow (B \Box\rightarrow \perp))$$

Similarly, the B schema  $\vdash A \rightarrow \Box\Diamond A$  is equivalent to this schema:

$$BS \vdash (A \Box\rightarrow (B \Box\rightarrow \perp)) \rightarrow (B \rightarrow (A \Box\rightarrow \perp))$$

The proofs are similar to some already given. The observations in this appendix merely begin the work of exploring the modal subsystems of logics of the counterfactual conditional. With luck, they will encourage others to explore the matter more thoroughly.

# Appendix 2

## Counterfactual Donkeys

---

This appendix experiments with an alternative way of formalizing the anaphora in the major premises of the arguments underlying philosophical thought experiments, by permitting the conditional in them to bind variables. Thus we formalize (5), (6) and (13) in Chapter 6 respectively as:

- (A1)  $GC(x, p) \square \rightarrow_{x,p} (JTB(x, p) \ \& \ \neg K(x, p))$
- (A2)  $(Farmer(x) \ \& \ Donkey(y) \ \& \ Owns(x, y)) \square \rightarrow_{x,p} Beats(x, y)$
- (A3)  $(Animal(x) \ \& \ Escapedzoo(x)) \square \rightarrow_x Monkey(x)$

These give a less unnatural treatment of the anaphora in (5), (6) and (13) than (3\*), (12) and (15) do, without repeating material or pulling a universal quantifier out of a hat. Of course, much depends on the semantics of this variable-binding conditional.

The natural strategy is to build on the preferred semantics for the conditional without variable-binding. Suppose, for a simple example, that we have a crude version of possible worlds semantics for  $\square \rightarrow$  (allowing, as usual, for vacuous truth):

- [ $\square \rightarrow$ ] A  $\square \rightarrow$  B is true at a world  $w$  if and only if B is true at the most similar worlds (if any) to  $w$  at which A is true.

Now think of assignments as assigning values both to the explicit variables and to a tacit world variable. Define an assignment  $s^*$  to be an  $x, \dots, y, w$ -variant of an assignment  $s$  just in case  $s^*$  differs from  $s$  at most over the values of the explicit variables  $x, \dots, y$  and of the world variable  $w$ . Then the modified semantic clause is this:

$[\Box \rightarrow_{x, \dots, y}]$   $A \Box \rightarrow_{x, \dots, y} B$  is true under an assignment  $s$  if and only if  
 $B$  is true at the most similar  $x, \dots, y, w$ -variants of  $s$   
(if any) to  $s$  at which  $A$  is true.

In effect,  $[\Box \rightarrow_{x, \dots, y}]$  replaces comparative similarity of worlds in  $[\Box \rightarrow]$  with comparative similarity of assignments, conceived as something like cases. Evidently, many more refined semantic clauses for  $\Box \rightarrow$  could be modified in corresponding ways.

Clause  $[\Box \rightarrow_{x, \dots, y}]$  corresponds to semantic clauses for variable-binding possibility and necessity operators:

$[\Diamond_{x, \dots, y}]$   $\Diamond_{x, \dots, y} A$  is true under an assignment  $s$  if and only if  $A$  is true at some  $x, \dots, y, w$ -variant of  $s$ .

$[\Box_{x, \dots, y}]$   $\Box_{x, \dots, y} A$  is true under an assignment  $s$  if and only if  $A$  is true at every  $x, \dots, y, w$ -variant of  $s$ .

The target analysis is expressible in this notation:

$$(A4) \quad \Box_{x,p} (K(x, p) \equiv JTB(x, p))$$

We could then rework the Gettier argument from (2) and (3\*) to (4) as an argument from (A1) and (A5) to (A6):

$$(A5) \quad \Diamond_{x,p} GC(x, p)$$

$$(A6) \quad \Diamond_{x,p} (JTB(x, p) \ \& \ \neg K(x, p))$$

Together,  $[\Box \rightarrow_{x, \dots, y}]$  and  $[\Diamond_{x, \dots, y}]$  validate the required analogue of the POSSIBILITY principle. The conclusion (A6) is inconsistent with the target analysis (A4), as expected.

Here is one advantage of formalizing (5) as (A1), understood in terms of a semantic clause like  $[\Box \rightarrow_{x, \dots, y}]$ , rather than as (3\*), understood in terms of a semantic clause like  $[\Box \rightarrow]$ . As noted in Chapter 6, section 5, (3\*) may be false in unexpected ways. For example, suppose that the Gettier case has many instances in the actual world; most of them are instances of justified true belief without knowledge, but a few abnormal instances are not instances of justified belief. Then (3\*) is false, because its antecedent is true and its consequent false in the actual world, even though most actual instances of the Gettier case are genuine counterexamples to the target analysis. In

such circumstances, is Gettier's argument really unsound? By contrast, (A1) understood in terms of  $[\Box \rightarrow_{x, \dots, y}]$  may avoid this problem, because the assignments which correspond to the normal instances may be closer to the assignment  $s$  with which we started than are the assignments which correspond to the abnormal instances. It is not implausible that they would be if we started with an assignment of ordinary objects to the explicit variables and the actual world to the world variable. Even abnormal instances in the actual world may be trumped in the overall similarity ranking by more ordinary realizations in counterfactual worlds. Although we could achieve some of the same effect by reading the quantifiers in (3\*) as contextually restricted, speakers may not know in advance how much restriction is needed, whereas  $[\Box \rightarrow_{x, \dots, y}]$  does not require any restriction to be specified in advance, and permits a flexible trade-off between similarity in the values of explicit variables and similarity in the value of the world variable.

The preceding remarks highlight an unusual feature of  $[\Box \rightarrow_{x, \dots, y}]$  as a semantic clause. Normally, the semantic clause for an operator  $O$  binding explicit variables has the effect that a closed formula with  $O$  as its main operator is true under all assignments if it is true under any. For example, on standard clauses for quantifiers, the truth-value of a closed quantified formula is independent of the assignment. Thus  $\forall x \forall x A$  and  $\exists x \forall x A$  are equivalent to  $\forall x A$  in any nonempty domain. By contrast, even when  $x, \dots, y$  exhaust the variables in  $A \Box \rightarrow_{x, \dots, y} B$ ,  $[\Box \rightarrow_{x, \dots, y}]$  allows it to be true under some assignments and false under others. In this it behaves with respect to both explicit variables and the world variable as counterfactual conditionals do with respect to the world variable in the absence of explicit variable-binding: a semantic clause like  $[\Box \rightarrow]$  allows  $A \Box \rightarrow B$  to be true at some worlds and false at others. Similarly,  $\Box \Box A$  and  $\Diamond \Box A$  are not equivalent to  $\Box A$  in many modal logics. Although the sensitivity of  $A \Box \rightarrow_{x, \dots, y} B$  in truth-value to the initial values of  $x, \dots, y$  creates no purely technical problem, it does raise the question where the explicit variables are to get their default values from. The context of utterance smoothly provides the world of the context as the default value of the world variable, but how is it to provide corresponding default values for the explicit variables?

It is in any case unlikely that a variable-binding operator in the object-language will give us everything we want, since the anaphora

in formulations of the Gettier argument can run across sentences, as in this wooden dialogue:

John: A person and a proposition could have stood in the Gettier relation.

Mary: If they had, they would have been an instance of justified true belief without knowledge.

The interaction of anaphora with intensional contexts creates notoriously thorny problems, which we obviously cannot attempt to solve here (for some of the issues see Roberts 1996). They do not show that the arguments underlying philosophical thought experiments are not to be understood in terms of counterfactual conditionals such as (5). They reveal some subtle obstacles to articulating a perfectly faithful formal analysis of those arguments, but for all they show the argument from (A1) and (A5) to (A6) (or from (2) and (3\*) to (4)) is a perfectly adequate approximation for almost all metaphysical purposes.

# Bibliography

---

Adams, E. W. 1965. “The logic of conditionals,” *Inquiry* 8: 166–97.

Adams, E. W. 1975. *The Logic of Conditionals*. Dordrecht: Reidel.

Anderson, A. R. 1951. “A note on subjunctive and counterfactual conditionals,” *Analysis* 12: 35–8.

Ayer, A. J. 1936. *Language, Truth and Logic*. London: Victor Gollancz.

Ayer, A. J. 1956. *The Problem of Knowledge*. London: Macmillan.

Bach, K. 1988. “Burge’s new thought experiment: back to the drawing room,” *Journal of Philosophy* 85: 88–97.

Bealer, G. 1998. “Intuition and the autonomy of philosophy,” in DePaul and Ramsey 1998.

Bealer, G. 2002. “Modal epistemology and the rationalist renaissance,” in Gendler and Hawthorne 2002.

Bennett, J. 2003. *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press.

Bergmann, G. 1964. *Logic and Reality*. Madison: The University of Wisconsin Press.

Bird, A. 1998. “Dispositions and antidotes,” *Philosophical Quarterly* 48: 227–34.

Blackburn, S. 1987. “Morals and modals,” in G. Macdonald and C. Wright, eds., *Fact, Science and Morality*. Oxford: Blackwell.

Boghossian, P. A. 1997. “Analyticity,” in R. Hale and C. Wright, eds., *A Companion to the Philosophy of Language*. Oxford: Blackwell.

Boghossian, P. A. 2002. “How are objective epistemic reasons possible?,” in J. Bermudez and A. Miller, eds., *Reason and Nature*. Oxford: Oxford University Press.

Boghossian, P. A. 2003. “Blind reasoning,” *The Aristotelian Society* sup. 77: 225–48.

Bonini, N., Osherson, D., Viale, R., and Williamson, T. 1999. “On the psychology of vague predicates,” *Mind & Language* 14: 377–93.

Braine, M. D., and O’Brien, D. P. 1991. “A theory of *if*: a lexical entry, reasoning program, and pragmatic principles,” *Psychological Review* 98: 182–203.

Brandom, R. 1994. *Making it Explicit: Reasoning, Representing, and Discursive Commitment*. Cambridge, MA: Harvard University Press.

Brandom, R. 2000. *Articulating Reasons: An Introduction to Inferentialism*. Cambridge, MA: Harvard University Press.

Burge, T. 1978. "Belief and synonymy," *Journal of Philosophy* 75: 119–38.

Burge, T. 1979. "Individualism and the mental," *Midwest Studies in Philosophy* 4: 73–121.

Burge, T. 1986. "Intellectual norms and foundations of mind," *Journal of Philosophy* 83: 697–720.

Byrne, R. M. J. 1989. "Suppressing valid inferences with conditionals," *Cognition* 31: 1–21.

Byrne, R. M. J. 2005. *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT Press.

Carnap, R. 1947. *Meaning and Necessity: A Study in Semantics and Modal Logic*. Chicago: The University of Chicago Press.

Carnap, R. 1950. *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Chalmers, D. J. 2002. "Does conceivability entail possibility?", in Gendler and Hawthorne 2002.

Chalmers, D. J. 2006. "The foundations of two-dimensional semantics," in M. García-Carpintero and J. Macià, eds., *Two-Dimensional Semantics*. Oxford: Clarendon Press.

Chambers, T. 1998. "On vagueness, *sorites*, and Putnam's 'intuitionistic strategy,'" *Monist* 81: 343–8.

Chisholm, R. 1957. *Perceiving: A Philosophical Study*. Ithaca, NY: Cornell University Press.

Cohen, S. 1988. "How to be a fallibilist," *Philosophical Perspectives* 2: 91–123.

Collins, J., Hall, N., and Paul, L. A., eds. 2004. *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Craig, E. 1985. "Arithmetic and fact," in I. Hacking, ed., *Essays in Analysis*. Cambridge: Cambridge University Press.

Cummins, R. 1998. "Reflections on reflective equilibrium," in DePaul and Ramsey 1998.

Currie, G. 1995a. *Image and Mind: Philosophy, Film and Cognitive Science*. New York: Cambridge University Press.

Currie, G. 1995b. "Visual imagery as the simulation of vision," *Mind and Language* 10: 17–44.

Davidson, D. 1974. "On the very idea of a conceptual scheme," *Proceedings and Addresses of the American Philosophical Association* 47: 5–20. Reprinted in Davidson 1984, to which page numbers refer.

Davidson, D. 1975. "Thought and talk," in S. Guttenplan, ed., *Mind and Language*, Oxford: Oxford University Press. Reprinted in Davidson 1984, to which page numbers refer.

Davidson, D. 1977. "The method of truth in metaphysics," in P. French, T. Uehling, and H. Wettstein, eds., *Midwest Studies in Philosophy*, 2: *Studies in the Philosophy of Language*. Morris: The University of Minnesota. Reprinted in Davidson 1984, to which page numbers refer.

Davidson, D. 1983. "A coherence theory of truth and knowledge," in D. Henrich, ed., *Kant oder Hegel?*. Stuttgart: Klett-Cotta. Reprinted with "Afterthoughts" in Davidson 2001.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press.

Davidson, D. 1986. "A nice derangement of epitaphs," in E. LePore, ed., *Truth and Interpretation*. Oxford: Blackwell.

Davidson, D. 1991. "Epistemology externalized," *Dialectica* 45: 191–202. Reprinted in Davidson 2001, to which page numbers refer.

Davidson, D. 2001. *Subjective, Intersubjective, Objective*. Oxford: Clarendon Press.

Davies, M. 1981. *Meaning, Quantification, Necessity*. London: Routledge and Kegan Paul.

Davies, M., and Humberstone, I. L. 1980. "Two notions of necessity," *Philosophical Studies* 38: 1–30.

Davies, M., and Stone, T., eds. 1995. *Mental Simulation: Evaluation and Applications*. Oxford: Blackwell.

DePaul, M. 1998. "Why bother with reflective equilibrium?," in DePaul and Ramsey 1998.

DePaul, M., and Ramsey, W., eds. 1998. *Rethinking Intuition: The Psychology of Intuition and its Role in Philosophical Inquiry*, Lanham, MD: Rowman and Littlefield.

DeRose, K. 1995. "Solving the skeptical problem," *Philosophical Review* 104: 1–52.

Dretske, F. 1970. "Epistemic operators," *Journal of Philosophy* 67: 1007–23.

Dummett, M. A. E. 1973. *Frege: Philosophy of Language*. London: Duckworth.

Dummett, M. A. E. 1975a. "Wang's paradox," *Synthese* 30: 301–24.

Dummett, M. A. E. 1975b. "The philosophical basis of intuitionistic logic," in H. Rose and J. C. Shepherdson, eds., *Logic Colloquium '73*. Amsterdam: North-Holland.

Dummett, M. A. E. 1977. *Elements of Intuitionism*. Oxford: Oxford University Press.

Dummett, M. A. E. 1978. *Truth and Other Enigmas*. London: Duckworth.

Dummett, M. A. E. 1993. *Origins of Analytical Philosophy*. London: Duckworth.

Edgington, D. 2001. "Conditionals," in L. Goble, ed., *The Blackwell Guide to Philosophical Logic*. Oxford: Blackwell.

Edgington, D. 2003. "Counterfactuals and the benefit of hindsight," in P. Dowe and P. Noordhof, eds., *Causation and Counterfactuals*. London: Routledge.

Eklund, M. 2002. "Inconsistent languages," *Philosophy and Phenomenological Research* 64: 251–75.

Elbourne, P. D. 2005. *Situations and Individuals*. Cambridge, MA: MIT Press.

Elugardo, R. 1993. "Burge on content," *Philosophy and Phenomenological Research* 53: 367–84.

Etchemendy, J. 1990. *The Concept of Logical Consequence*. Cambridge, MA: Cambridge University Press.

Evans, G. 1979. "Reference and contingency," *Monist* 62: 161–89.

Evans, G. 1982. *The Varieties of Reference*. Oxford: Clarendon Press.

Evans, G. 1985. *Collected Papers*. Oxford: Clarendon Press.

Evans, J. St. B. T., and Over, D. E. 2004. *If*. Oxford: Oxford University Press.

Feyerabend, P. 1978. *Science in a Free Society*. London: NLB.

Fine, K. 1970. "Propositional quantifiers in modal logic," *Theoria* 36: 336–46.

Fine, K. 1994. "Essence and modality," in J. Tomberlin, ed., *Philosophical Perspectives*, 8: *Logic and Language*. Atascadero, CA: Ridgeview.

Fine, K. 1995. "Senses of essence," in W. Sinnott-Armstrong, D. Raffman, and N. Asher, eds., *Modality, Morality, and Belief: Essays in Honor of Ruth Barcan Marcus*. Cambridge: Cambridge University Press.

Fodor, J. A. 1975. *The Language of Thought*. New York: Thomas Crowell.

Fodor, J. A. 1998. *Concepts: Where Cognitive Science Went Wrong*. Oxford: Oxford University Press.

Føllesdal, D. 2004. *Referential Opacity and Modal Logic*. London: Routledge.

Frege, G. 1950. *The Foundations of Arithmetic*, trans. J. L. Austin. Oxford: Blackwell.

French, P. A., Uehling, T. E., and Wettstein, H. K., eds. 1986. *Midwest Studies in Philosophy XI: Studies in Essentialism*. Minneapolis: University of Minnesota Press.

Gallie, W. B. 1964. *Philosophy and the Historical Understanding*. London: Chatto & Windus.

Gauker, C. 2005. *Conditionals in Context*. Cambridge, MA: MIT Press.

Gaut, B. 2006. "Art and Cognition," in M. Kieran, ed., *Contemporary Debates in Aesthetics and the Philosophy of Art*. Oxford: Blackwell.

Gendler, T. S. 1998. "Galileo and the indispensability of scientific thought experiments," *British Journal for the Philosophy of Science* 49: 397–424.

Gendler, T. S. 2004. "Thoughts experiments rethought – and reperceived," *Philosophy of Science* 71: 1152–63.

Gendler, T. S., and Hawthorne, J., eds. 2002. *Conceivability and Possibility*. Oxford: Clarendon Press.

Gettier, E. 1963. "Is justified true belief knowledge?," *Analysis* 23: 121–3.

Goldberg, S. 2000. "Do anti-individualistic construals of propositional attitudes capture the agent's conceptions?," *Noûs* 36: 597–621.

Goldman, A. 1992. "Empathy, mind, and morals," *Proceedings and Addresses of the American Philosophical Association* 66/3: 17–41.

Goldman, A. 2005. "Kornblith's naturalistic epistemology," *Philosophy and Phenomenological Research* 71: 403–10.

Goldman, A., and Pust, J. 1998. "Philosophical theory and intuitional evidence," in DePaul and Ramsey 1998.

Goodman, N. 1955. *Fact, Fiction and Forecast*. Cambridge, MA: Harvard University Press.

Graff, D., and Williamson, T., eds. 2002. *Vagueness*. Aldershot: Dartmouth.

Grandy, R. 1973. "Reference, meaning, and belief," *Journal of Philosophy* 70: 439–52.

Grice, H. P. 1961. "The causal theory of perception," *Proceedings of the Aristotelian Society* sup. 35: 121–52.

Grice, H. P. 1986. "Reply to Richards," in R. Grandy and R. Warner, eds., *Philosophical Grounds of Rationality: Intentions, Categories, Ends*. Oxford: Clarendon Press.

Grice, H. P. 1989. *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Grice, H. P., and Strawson, P. F. 1956. "In defence of a dogma," *Philosophical Review* 65: 141–58.

Häggqvist, S. 1996. *Thought Experiments in Philosophy*. Stockholm: Almqvist and Wiksell.

Halbach, V. 2001. "How innocent is deflationism?," *Synthese* 126: 167–94.

Harman, G. 1986. *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press.

Harman, G. 1999. *Reasoning, Meaning, and Mind*. Oxford: Clarendon Press.

Harris, P. 2000. *The Work of the Imagination*. Oxford: Blackwell.

Hawthorne, J. 2004. *Knowledge and Lotteries*. Oxford: Oxford University Press.

Hawthorne, J. 2006. *Metaphysical Essays*. Oxford: Clarendon Press.

Hempel, C. G. 1965. *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. New York: The Free Press.

Hill, C. S. 2006. "Modality, modal epistemology, and the metaphysics of consciousness," in S. Nichols (ed.), *The Architecture of the Imagination: New Essays on Pretense, Possibility and Fiction*. Oxford: Oxford University Press.

Hintikka, J. 1999. "The Emperor's new intuitions," *Journal of Philosophy* 96: 127–47.

Horgan, T. 1995. "Transvaluationism: a Dionysian approach to vagueness," *Southern Journal of Philosophy* 33, sup.: 97–126.

Horgan, T. 1998. "The transvaluationist conception of vagueness," *The Monist* 81: 313–30.

Horwich, P. 1998. *Meaning*. Oxford: Clarendon Press.

Hughes, G., and Cresswell, M. 1996. *A New Introduction to Modal Logic*. London: Routledge.

Jackson, F. 1977. "A causal theory of counterfactuals," *Australasian Journal of Philosophy* 55: 3–21.

Jackson, F. 1979. "On assertion and indicative conditionals," *Philosophical Review* 88: 565–89.

Jackson, F. 1981. "Conditionals and possibilia," *Proceedings of the Aristotelian Society* 81: 125–37.

Jackson, F. 1987. *Conditionals*. Oxford: Blackwell.

Jackson, F. 1998. *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Clarendon Press.

Jackson, F. 2001. "Responses," *Philosophy and Phenomenological Research*, 62: 653–64.

Johnson-Laird, P. N., and Byrne, R. M. J. 1993. "Models and deductive rationality," in K. Manktelow and D. Over, eds., *Rationality: Psychological and Philosophical Perspectives*. London: Routledge.

Johnston, M. 1993. "Objectivity refigured: pragmatism without verificationism," in J. Haldane and C. Wright, eds., *Reality, Representation and Projection*. Oxford: Oxford University Press.

Jönsson, M. L., and Hampton, J. A. 2006. "The inverse conjunction fallacy," *Journal of Memory and Language* 55: 317–34.

Kahneman, D., and Frederick, S. 2002. "Representativeness revisited: attribute substitution in intuitive judgment," in T. Gilovich, D. Griffin, and D. Kahneman, eds., *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.

Kahneman, D., and Tversky, A. 1982. "The simulation heuristic," in D. Kahneman, P. Slovic, and A. Tversky, eds., *Judgement under Uncertainty*. Cambridge: Cambridge University Press.

Kaplan, D. 1970. "S5 with quantifiable propositional variables," *Journal of Symbolic Logic* 35: 355.

Kaplan, D. 1989. "Demonstratives: an essay on the semantics, logic metaphysics, and epistemology of demonstratives and other indexicals," in J. Almog, J. Perry, and H. Wettstein, eds., *Themes from Kaplan*. Oxford: Oxford University Press.

Kaplan, D. 1990. "Words," *Aristotelian Society* sup. 64: 93–119.

Keefe, R. 2000. *Theories of Vagueness*. Cambridge: Cambridge University Press.

Keefe, R., and Smith, P., eds. 1997. *Vagueness: A Reader*. Cambridge, MA: MIT Press.

Kleene, S. C. 1952. *Introduction to Metamathematics*. Amsterdam: North-Holland.

Klein, P. 1986. "Radical interpretation and global skepticism," in LePore 1986.

Kment, B. 2006. "Counterfactuals and the analysis of necessity," *Philosophical Perspectives* 20: 237–302.

Kornblith, H. 2002. *Knowledge and its Place in Nature*. Oxford: Oxford University Press.

Kornblith, H. 2006. "Appeals to intuition and the ambitions of epistemology," in S. Hetherington, ed., *Epistemology Futures*. Oxford: Clarendon Press.

Kornblith, H. 2007. "Naturalism and intuitions," *Grazer Philosophische Studien* 74: 27–49.

Kratzer, A. 1977. "What 'must' and 'can' must and can mean," *Linguistics and Philosophy* 1: 337–55.

Kratzer, A. 1986. "Conditionals," in A. M. Farley, P. Farley, and K. E. McCollough, eds., *Papers from the Parasession on Pragmatics and Grammatical Theory*. Chicago: Chicago Linguistics Society.

Kripke, S. A. 1979. "A puzzle about belief," in A. Margalit, ed., *Meaning and Use*. Dordrecht: Reidel.

Kripke, S. A. 1980. *Naming and Necessity*. Oxford: Blackwell.

Ladusaw, W. 1996. "Negation and polarity items," in S. Lappin, ed., *The Handbook of Contemporary Semantic Theory*. Oxford: Blackwell.

Lange, M. 2005. "A counterfactual analysis of logical truth and necessity," *Philosophical Studies* 125: 277–303.

Langford, C. H. 1942. "The notion of analysis in Moore's philosophy," in P. A. Schilpp, ed., *The Philosophy of G. E. Moore*. Evanston: Northwestern University Press.

LePore, E., ed. 1986. *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Oxford: Blackwell.

Lewis, D. K. 1973a. "Counterfactuals and comparative possibility," *Journal of Philosophical Logic* 2: 418–46. Reprinted in his *Philosophical Papers*, vol. 2. Oxford: Oxford University Press, 1986, to which page numbers refer.

Lewis, D. K. 1973b. "Causation," *Journal of Philosophy* 70: 556–67.

Lewis, D. K. 1974. "Radical interpretation," *Synthese* 23: 331–44. Reprinted with "Postscripts" in his *Philosophical Papers*, vol. 1. Oxford: Oxford University Press, 1983, to which pages number refer.

Lewis, D. K. 1975. "Adverbs of quantification," in E. Keenan, ed., *Formal Semantics of Natural Language*. Cambridge: Cambridge University Press.

Lewis, D. K. 1979. "Counterfactual dependence and time's arrow," *Noûs* 13: 455–76.

Lewis, D. K. 1983a. *Philosophical Papers*, vol. 1. Oxford: Oxford University Press.

Lewis, D. K. 1983b. "New work for a theory of universals," *Australasian Journal of Philosophy* 61: 343–77.

Lewis, D. K. 1986. *Counterfactuals*, revised edn. Cambridge, MA: Harvard University Press.

Lewis, D. K. 1996. "Elusive knowledge," *Australasian Journal of Philosophy* 74: 549–67.

Lewis, D. K. 1997. "Finkish dispositions," *Philosophical Quarterly* 47: 143–58.

Lowe, E. J. 1987. "Not a counterexample to modus ponens," *Analysis* 47: 44–7.

Lycan, W. G. 2001. *Real Conditionals*. Oxford: Oxford University Press.

Manktelow, K. I., and Over, D. E. 1987. "Reasoning and rationality," *Mind and Language* 2: 199–219.

Marconi, D. 1997. *Lexical Competence*. Cambridge, MA: MIT Press.

Marion, M. 2000. "Oxford realism: knowledge and perception," parts I and II, *British Journal for the History of Philosophy* 8: 299–338, 485–519.

Martin, C. B. 1994. "Dispositions and conditionals," *Philosophical Quarterly* 44: 1–8.

Martin, C. B., and Heil, J. 1998. "Rules and powers," *Philosophical Perspectives* 12: 283–312.

Mates, B. 1952. "Synonymity," in L. Linsky, ed., *Semantics and the Philosophy of Language*. Urbana: University of Illinois Press.

McCulloch, G. 2003. *The Life of the Mind: An Essay on Phenomenological Externalism*. London: Routledge.

McDowell, J. H. 1994. *Mind and World*. Cambridge, MA: Harvard University Press.

McGee, V. 1985. "A counterexample to modus ponens," *Journal of Philosophy* 82: 462–71.

McGee, V., and McLaughlin, B. 2000. "The lessons of the many," *Philosophical Topics* 28: 129–51.

McGinn, C. 1986. "Radical interpretation and epistemology," in LePore 1986.

McKinsey, M. 1991. "Anti-individualism and privileged access," *Analysis* 51: 9–16.

Moore, G. E. 1925. "A defence of common sense," in J. Muirhead, ed., *Contemporary British Philosophy* (2nd series). London: George Allen & Unwin.

Mumford, S. 1998. *Dispositions*. Oxford: Oxford University Press.

Nagel, T. 1986. *The View from Nowhere*. Oxford: Oxford University Press.

Neale, S. 1990. *Descriptions*. Cambridge, MA: MIT Press.

Newstead, S.E., Handley, S.J., Harley, C., Wright, H., and Farrelly, D. 2004. "Individual differences in deductive reasoning," *Quarterly Journal of Experimental Psychology* 57A: 33–60.

Nichols, S., and Stich, S. P. 2003. *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds*. Oxford: Clarendon Press.

Nichols, S., Stich, S. P., Leslie, A., and Klein, D. 1996. "Varieties of off-line simulation," in P. Carruthers and P. K. Smith, eds., *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

Nolan, D. 1997. "Impossible worlds: a modest approach," *Notre Dame Journal for Formal Logic* 38: 535–72.

Nolan, D. 2003. "Defending a possible-worlds account of indicative conditionals," *Philosophical Studies* 116: 215–69.

Norton, J. D. 1991. "Thought experiments in Einstein's work," in T. Horowitz and G. J. Massey, eds., *Thought Experiments in Science and Philosophy*. Savage, MD: Rowman and Littlefield.

Norton, J. D. 2004. "Why thought experiments do not transcend empiricism," in C. Hitchcock, ed., *Contemporary Debates in the Philosophy of Science*. Oxford: Blackwell.

Nozick, R. 1981. *Philosophical Explanations*. Oxford: Oxford University Press.

Nozick, R. 2001. *Invariances: The Structure of the Objective World*. Cambridge, MA: Harvard University Press.

Oaksford, M. 2005. "Reasoning," in N. Braisby and M. Gellatly, eds., *Cognitive Psychology*. Oxford: Oxford University Press.

Over, D. E. 1987. "Assumptions and the supposed counterexamples to modus ponens," *Analysis* 47: 142–6.

Peacocke, C. A. B. 1985. "Imagination, experience and possibility," in J. Foster and H. Robinson, eds., *Essays on Berkeley: A Tercentennial Celebration*. Oxford: Clarendon Press.

Peacocke, C. A. B. 1992. *A Study of Concepts*. Cambridge, MA: MIT Press.

Peacocke, C. A. B. 1999. *Being Known*. Oxford: Clarendon Press.

Peacocke, C. A. B. 2004. *The Realm of Reason*. Oxford: Oxford University Press.

Priest, G. 1995. *Beyond the Limits of Thought*. Cambridge: Cambridge University Press.

Priest, G., Beall, J.C., and Armour-Garb, B., eds. 2004. *The Law of Non-Contradiction: New Philosophical Essays*. Oxford: Clarendon Press.

Prior, A. N. 1960. "The runabout inference-ticket," *Analysis* 21: 38–9.

Pust, J. 2001. "Against explanationist skepticism regarding philosophical intuitions," *Philosophical Studies* 106: 227–58.

Putnam, H. 1975. *Mind, Language and Reality: Philosophical Papers, Volume 2*. Cambridge: Cambridge University Press.

Quine, W. V. O. 1936. "Truth by convention," in O. H. Lee, ed., *Philosophical Essays for A. N. Whitehead*. New York: Longmans.

Quine, W. V. O. 1951. "Two dogmas of empiricism," *Philosophical Review* 60: 20–43.

Quine, W. V. O. 1953. *From a Logical Point of View*. Cambridge, MA: Harvard University Press.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: MIT Press.

Quine, W. V. O. 1966. *The Ways of Paradox and Other Essays*. New York: Random House.

Quine, W. V. O. 1970. *Philosophy of Logic*. Englewood Cliffs, NJ: Prentice-Hall.

Ramsey, F. P. 1978. *Foundations: Essays in Philosophy, Logic, Mathematics and Economics*, ed. D. H. Mellor. London: Routledge & Kegan Paul.

Rawls, J. 1951. "Outline of a decision procedure for ethics," *Philosophical Review* 60: 167–97.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Roberts, C. 1996. "Anaphora in intensional contexts," in S. Lappin, ed., *The Handbook of Contemporary Semantic Theory*. Oxford: Blackwell.

Roese, N. J., and Olson, J. 1993. "The structure of counterfactual thought," *Personality and Social Psychology Bulletin* 19: 312–19.

Roese, N. J., and Olson, J. 1995. "Functions of counterfactual thinking," in N. J. Roese and J. M. Olson, eds., *What Might Have Been: The Social Psychology of Counterfactual Thinking*. Mahwah, NJ: Erlbaum.

Rorty, R., ed. 1967. *The Linguistic Turn: Recent Essays in Philosophical Method*. Chicago: University of Chicago Press.

Rosen, G. 1990. "Modal fictionalism," *Mind* 99: 327–54.

Russell, B. A. W. 1912. *The Problems of Philosophy*. London: Williams and Norgate.

Sainsbury, R. M. 1997. "Easy possibilities," *Philosophy and Phenomenological Research* 57: 907–19.

Salmon, N. 1982. *Reference and Essence*. Oxford: Blackwell.

Salmon, N. 1986. *Frege's Puzzle*. Cambridge, MA: MIT Press.

Salmon, N. 1989. "The logic of what might have been," *Philosophical Review* 98: 3–34.

Salmon, N. 1993. "This side of paradox," *Philosophical Topics* 21: 187–97.

Schechter, J. 2006. "Can evolution explain the reliability of our logical beliefs," typescript.

Schroeter, L., and Schroeter, F. 2006. "Rational improvisation," typescript.

Schroyens, W., and Schaeken, W. 2003. "A critique of Oaksford, Chater, and Larkin's (2000) conditional probability model of conditional reasoning," *Journal of Experimental Psychology: Learning, Memory and Cognition* 29: 140–9.

Shope, R. 1983. *The Analysis of Knowing: A Decade of Research*. Princeton: Princeton University Press.

Sides, A., Osherson, D., Bonini, N., and Viale, R. 2002. "On the reality of the conjunction fallacy," *Memory and Cognition* 30: 191–8.

Sinnott-Armstrong, W. 1999. "Begging the question," *Australasian Journal of Philosophy* 77: 174–91.

Sinnott-Armstrong, W., Moor, J., and Fogelin, R. 1986. "A defense of modus ponens," *Journal of Philosophy* 83: 296–300.

Smart, J. J. C. 1984. *Ethics, Persuasion and Truth*. London: Routledge, Kegan & Paul.

Smart, J. J. C. 1987. *Essays Metaphysical and Moral: Selected Philosophical Papers*. Oxford: Blackwell.

Soames, S. 1995. "T-sentences," in W. Sinnott-Armstrong, D. Raffman, and N. Asher, eds., *Modality, Morality and Belief: Essays in Honor of Ruth Barcan Marcus*. Cambridge: Cambridge University Press.

Soames, S. 1999. *Understanding Truth*. Oxford: Oxford University Press.

Sober, E. 2000. "Quine," *Aristotelian Society* sup. 74: 237–80.

Sorensen, R. A. 1992. *Thought Experiments*. Oxford: Oxford University Press.

Sorensen, R. A. 2001. *Vagueness and Contradiction*. Oxford: Clarendon Press.

Sosa, E. 2005. "A defense of the use of intuitions in philosophy," in M. Bishop and D. Murphy, eds., *Stich and His Critics*. Oxford: Blackwell.

Sosa, E. 2006. "Intuitions and truth," in P. Greenough and M. P. Lynch, eds., *Truth and Realism*. Oxford: Clarendon Press.

Stalnaker, R. C. 1968. "A theory of conditionals," in *American Philosophical Quarterly Monographs 2 (Studies in Logical Theory)*: 98–112.

Stalnaker, R. C. 1984. *Inquiry*. Cambridge, MA: MIT Press.

Stalnaker, R. C. 1999. *Context and Content*. Oxford: Oxford University Press.

Stalnaker, R. C. 2003. *Ways a World Might Be*. Oxford: Clarendon Press.

Stanley, J. 2005. *Knowledge and Practical Interests*, Oxford: Oxford University Press.

Stanovich, K. E., and West, R. F. 2000. "Individual differences in reasoning: implications for the rationality debate?," *Behavioral and Brain Sciences* 23: 645–65.

Stich, S. 1998. "Reflective equilibrium, analytic epistemology and the problem of cognitive diversity," in DePaul and Ramsey 1998.

Stine, G. 1976. "Skepticism, relevant alternatives, and deductive closure," *Philosophical Studies* 29: 249–61.

Strawson, P. F. 1959. *Individuals: An Essay in Descriptive Metaphysics*. London: Methuen.

Sutton, J. 2007. *Without Justification*. Cambridge, MA: MIT Press.

Tappenden, J. 1993. "Analytic truth – it's worse (or perhaps better) than you thought," *Philosophical Topics* 21: 233–61.

Tappolet, C. 1997. "Mixed inferences: a problem for pluralism about truth predicates," *Analysis* 57: 209–10.

Tarski, A. 1983a. "The concept of truth in formalized languages," trans J. H. Woodger, in *Logic, Semantics, Metamathematics*, 2nd edn., J. Corcoran, ed. Indianapolis: Hackett.

Tarski, A. 1983b. "On the concept of logical consequence," trans. J. H. Woodger, in *Logic, Semantics, Metamathematics*, 2nd edn., J. Corcoran, ed. Indianapolis: Hackett.

Tversky, A., and Kahneman, D. 1983. "Extensional versus intuitive reasoning: the conjunction fallacy in probability judgment," *Psychological Review* 90: 293–315.

Vahid, H. 2004. "Varieties of epistemic conservativism," *Synthese* 141: 97–122.

van Inwagen, P. 1995. *Material Beings*. Ithaca, NY: Cornell University Press.

van Inwagen, P. 1997. "Materialism and the psychological-continuity account of personal identity," in J. Tomberlin, ed., *Philosophical Perspectives*, 11: *Mind, Causation and World*. Oxford: Blackwell.

van Rooij, R. 2006. "Free choice counterfactual donkeys," *Journal of Semantics* 23: 383–402.

von Fintel, K. 2001. "Counterfactuals in a dynamic context," in M. Kenstowicz, ed., *Ken Hale: A Life in Language*. Cambridge, MA: MIT Press.

Wason, P. C., and Shapiro, D. 1971. "Natural and contrived experience in a reasoning problem," *Quarterly Journal of Experimental Psychology* 23: 63–71.

Weatherson, B. 2003. "What good are counterexamples?," *Philosophical Studies* 115: 1–31.

Weinberg, J., Stich, S., and Nichols, S. 2001. "Normativity and epistemic intuitions," *Philosophical Topics* 29: 429–60.

Wiggins, D. R. P. 2001. *Sameness and Substance Renewed*. Cambridge: Cambridge University Press.

Williams, B. 1966. "Imagination and the self," *Proceedings of the British Academy* 52: 105–24.

Williamson, T. 1990. *Identity and Discrimination*. Oxford: Blackwell.

Williamson, T. 1994a. *Vagueness*. London: Routledge.

Williamson, T. 1994b. "Crispin Wright, *Truth and Objectivity*," *International Journal of Philosophical Studies* 2: 130–44.

Williamson, T. 1999a. "Truthmakers and the converse Barcan formula," *Dialectica* 53: 253–70.

Williamson, T. 1999b. "Schiffer on the epistemic theory of vagueness," *Philosophical Perspectives* 13: 505–17.

Williamson, T. 2000a. *Knowledge and its Limits*. Oxford: Oxford University Press.

Williamson, T. 2000b. "Existence and contingency," *Proceedings of the Aristotelian Society* 100: 117–39.

Williamson, T. 2001. "Ethics, supervenience and Ramsey sentences," *Philosophy and Phenomenological Research* 62: 625–30.

Williamson, T. 2003a. "Understanding and inference," *Aristotelian Society* sup. 77: 249–93.

Williamson, T. 2003b. "Vagueness in reality," in M. Loux and D. Zimmerman, eds., *The Oxford Handbook of Metaphysics*. Oxford: Oxford University Press.

Williamson, T. 2004a. "Philosophical 'intuitions' and scepticism about judgement," *Dialectica* 58: 109–53.

Williamson, T. 2004b. "Past the linguistic turn?," in B. Leiter, ed., *The Future for Philosophy*. Oxford: Oxford University Press.

Williamson, T. 2005a. "Armchair philosophy, metaphysical modality and counterfactual thinking," *Proceedings of the Aristotelian Society* 105: 1–23.

Williamson, T. 2005b. "Contextualism, subject-sensitive invariantism and knowledge of knowledge," *Philosophical Quarterly* 55: 213–35.

Williamson, T. 2005c. "Knowledge and scepticism," in F. Jackson and M. Smith, eds., *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press.

Williamson, T. 2006a. "Indicative versus subjunctive conditionals, congruential versus non-hyperintensional contexts," *Philosophical Issues* 16: 310–33.

Williamson, T. 2006b. "Conceptual truth," *Aristotelian Society* sup. 80: 1–41.

Williamson, T. 2006c. "Must do better," in P. Greenough and M. Lynch, eds., *Truth and Realism*. Oxford: Oxford University Press.

Williamson, T. 2007. "Philosophical knowledge and knowledge of counterfactuals," *Grazer Philosophische Studien* 74: 89–123.

Williamson, T. 2008a. "Why epistemology can't be operationalized," in Q. Smith, ed., *Epistemology: New Philosophical Essays*. Oxford: Oxford University Press.

Williamson, T. 2008b. "Reference, inference and the semantics of pejoratives," in J. Almog and P. Leonardi, eds., *Essays for David Kaplan*, Oxford: Oxford University Press.

Wilson, J. C. 1926. *Statement and Inference*, 2 vols. Oxford: Clarendon Press.

Wright, C. J. G. 1989. "Necessity, caution and scepticism," *Aristotelian Society* sup. 63: 203–38.

Wright, C. J. G. 1993. *Realism, Meaning and Truth*, 2nd edn. Oxford: Blackwell.

Yablo, S. 2002. "Coulda, woulda, shoulda," in Gendler and Hawthorne 2002.

Zimmerman, D. 2004. "Prologue: metaphysics after the twentieth century," *Oxford Studies in Metaphysics* 1: ix–xxii.

# Index

---

*a posteriori* methods 1–2, 61–2, 70, 280

*a priori* methods 1–2, 11, 25, 46, 48, 51, 62, 65, 69

*a priori / a posteriori* distinction 2–3, 136, 165–9, 189–90, 195

abduction *see* inference to the best explanation

action

- and belief 252–8
- and knowledge 269–70

“actually” operator 62, 64–5, 138, 144–5, 152–3, 295–6

aesthetic criteria 274–5, 285, 289

agreement 1, 7, 37–8, 41, 96–7, 114–15, 121, 125–6, 128, 190–2, 211, 213, 221–2, 246–8, 250, 260–2, 276, 280, 285, 287–8, 290

analysis 11–15, 18, 20, 46, 48, 69–70, 121n34, 141, 161n11, 178–85, 187, 193–4, 203, 205, 235n16, 243, 280, 306

analytic philosophy 2, 12–15, 18, 21, 45–6, 180, 214–15, 235–6, 279, 286, 288

analyticity 5, 11, 20–21, 25, 46–74, 77, 85, 92, 102, 109–10, 112, 116–19, 125–6, 130–2, 188, 218

*see also* epistemology of analytic truths; Frege-analyticity; metaphysics of analytic truths; quasi-Frege-analyticity; modal-analyticity

Anaximenes 278

Anderson, A. R. 137

Anscombe, G. E. M. 21

anthropocentrism 262

anti-individualism *see* externalism, semantic

appearance principles 227–31

appearances 216–17, 223, 225–32, 234–5

Aristotle 19, 269n11

armchair methods 1, 4, 6–7, 19, 21, 25, 45, 47–8, 52, 60, 73, 76–7, 84, 132, 169, 179, 279

*see also* *a priori* methods; *a priori / a posteriori* distinction

Armstrong, D. M. 19, 21

art, philosophy of 49

assent 74–5, 78, 113, 116–17, 129–30, 132

*see also* understanding–assent links

assertability-conditional semantics 281–5

astrology 241

Austin, J. L. 21, 270n11

Ayer, A. J. 11–12, 14, 21, 47, 180

B (modal logic) 304

Bach, K. 98n18

Bayesianism 139–40

Bealer, G. 74, 216, 226

beliefs and truth 247–8, 251–8, 260–77

Bennett, J. 152n8, 176

Bergmann, G. 10

Berkeley, G. 11, 149n6

biology 4, 18, 20, 221, 285

biology, philosophy of 4, 6, 18

Bird, A. 100n20

Blackburn, S. 136

Boghossian, P. 53n4, 59n9, 63n14, 69n21, 76, 80–2, 84n7, 88n10, 92, 95n16, 121n34

Bonini, N. 6n1, 96n17

Braine, M. 105n24

Brandom, R. 76, 80n4

Burge, T. 91, 97–8, 117, 124

Byrne, R. 103, 106, 140–1

Campbell, J. 13

Carnap, R. 21, 51, 121–2, 127, 239n18, 282

causal relations 5–6, 17, 21, 66, 72, 123–4, 126–7, 129, 140–1, 149, 166, 168, 203, 217, 220, 250, 257–9, 263–4, 266–7, 271

*see also* reference, causal theories of

centering axiom 152, 294

Chalmers, D. 169n13

Chambers, T. 32n6

charity, principles of 125, 260–77

*see also* knowledge maximization

Chisholm, R. 180

Chomsky, N. 122, 212n2

Churchland, P. M. 220

closure principles

for confidence 230, 234

for counterfactual conditionals 143–5, 154n9, 159–60, 187, 293–6

for justification 181–2

for knowledge 234

Cohen, S. 93n14

common sense *see* folk theories

competence / performance distinction 99–101

conceivability 135, 163

*see also* imagination

concepts 3, 13–18, 20, 29–30, 39, 41, 48, 52, 74–5, 77, 81–2, 114, 116–19, 129–30, 165–7, 183, 191, 206, 211, 220–1, 258, 280n1

conceptual competence 2–3, 39, 41, 48, 52–3, 73–4, 76–7, 81, 83–4, 95n16, 113–14, 116, 130–3, 166, 168, 189–90, 216, 218, 258–9

conceptual philosophy 2–4, 13–15, 18

conceptual possibility 205–7

conceptual questions 2–4, 6, 48

conceptual skills 166–9, 191, 216, 220

conceptual truth *see* analyticity

conceptual turn 2, 14, 18–19, 21–22, 31, 49, 53

conceptualism 19

conditionals 35n8, 56–8, 64, 82, 87–8, 92–5, 100, 103, 106–8, 137–65, 186, 195–9, 205–6, 305–8

*see also* counterfactual conditionals; modus ponens

conjunction elimination 95–6

conjunction fallacy 96

conjunction introduction 95  
 consequence fallacy 233–4  
 conservativism 7, 223, 242–3  
 constructivism 260  
 context sensitivity 29n5, 59–60,  
     62, 93, 114, 128, 140, 175–7,  
     200, 220, 223, 231, 268, 307  
 conventional implicatures 128–9  
 conversational maxims 96  
 cotenability 143–4, 150, 153  
 counterfactual conditionals  
     108n27, 137–65, 169–77,  
     186, 194–200, 204–5,  
     293–308  
 embedding capacity of 147n4,  
     161, 176–7  
 logic 143–5, 152–4, 171–5, 187,  
     293–304  
 with donkey anaphora 195–9,  
     305–8  
*see also* epistemology of  
     counterfactual conditionals;  
     logic of counterfactual  
     conditionals  
 counterpossibles 171–5  
 Craig, E. J. 136  
 Cresswell, M. 302  
 Cummins, R. 244n20  
 Currie, G. 149n6  
 Davidson, D. 21, 122, 125n38,  
     260–1, 265–6, 271–2, 282  
 Davies, M. 65, 199n17, 295  
 definitions 11–12, 20–21, 49–50,  
     54, 64n15, 71–2, 118, 121–2,  
     289  
 DePaul, M. 244n20  
 DeRose, K. 93n14, 234n14  
 Derrida, J. 12  
 Descartes, R. 19, 236  
 descriptivism 69, 77, 122n34,  
     263–4  
 desires and goodness 252–7  
 deviant logic 88, 91, 97, 108, 115,  
     273  
 dialectics 1, 136, 173–4, 182, 203,  
     210–11, 215, 238–40, 276  
 dialetheism 94–5, 126  
 direct reference 66–7, 69  
 disagreement *see* agreement  
 disjunctive syllogism 94  
 dispositions 99–102, 104–5, 125,  
     132, 136, 243, 283  
*see also* understanding–  
     disposition-to-assent links  
 disquotation 27–8, 54–6, 79, 83,  
     90, 110–11, 247n1, 280  
 division of linguistic labor 94, 98,  
     123  
 donkey sentences *see* counterfactual  
     conditionals with donkey  
     anaphora  
 Dretske, F. 234n14  
 Dummett, M. 11–15, 20–1, 32n6,  
     67n19, 76, 80n4, 120n33,  
     258n6, 281–4, 290  
 easy possibility *see* possibility,  
     easy  
 economics, philosophy of 6  
 Edgington, D. 93n15, 137, 176  
 Einstein, A. 179, 279  
 Eklund, M. 79n3, 100n20, 120n33  
 Elbourne, P. 196n15, 199n17  
 eliminativist theories 220, 237  
 Elugardo, R. 98n18  
 elusive objects 16–17  
 empiricism 1–2, 4, 11  
 empty terms *see* reference failure  
 enabling role of experience 165–9,  
     189–90  
 epistemic conservativism 242–3  
 epistemic possibility 207  
 epistemicism 43

epistemology 5, 12, 49, 51, 60, 93n14, 164, 180, 188–9, 193–4, 203, 206–7, 211, 216, 236, 244, 247, 249–50, 261–2, 271, 273

of analytic truths 53–4, 60, 62–3, 66–8, 70–3, 77, 81, 84, 130–1

of counterfactual conditionals 136, 138–55, 158, 162–80, 188–9, 216

of logic 53, 65–8, 70, 162–3, 171, 239–40, 245–6, 274

of metaphysical modality 53, 62, 134–7, 155, 158, 162–5, 167, 169–71, 178–80, 189

of philosophy 5, 50, 187–95, 206–7, 208–47, 250–1, 273–7

essentialism 19–20

about origin 161

essentially contested concepts 126n40

Etchemendy, J. 53n5

Evans, G. 13, 60n12, 111, 114n29

Evans, J. 105n24, 142

evidence 5, 42, 73, 81, 138–40, 165, 203, 208–14, 220, 222–8, 230, 232–42, 244–6, 274–7, 279

and probability 227–32

as propositional 208–9

psychologization of 5, 211–12, 214, 226, 234–8, 241, 243, 245–6, 274

as true 209–10

*see also* total evidence, requirement of

evidence neutrality 210–14, 234–7

evidential role of experience 165–9, 190

evolution 4, 126, 136, 190, 221, 251–2, 255–7

exceptionalism, philosophical 3–4, 6, 133, 136–7, 178–80, 188, 191–2, 208, 246, 277

excluded middle, law of 31–4, 44, 290

existential commitment 86, 88, 105, 173

experimental philosophy 6, 190–1

experiments 1, 6, 19, 21, 141–2, 179, 191, 208–10, 212, 276, 279, 287–8

*see also* thought experiments

externalism, epistemological 209, 271

externalism, semantic 91, 94, 97–8, 118–19, 123–5, 169n13

falsity-indicativeness 227–31

Feferman, S. 280

Feyerabend, P. 224n7

fiction 183–4, 192, 195, 216, 223

Fine, K. 19, 21, 296n1

Fodor, J. 13, 21, 70n23

Fogelin, R. 93n12

Føllesdal, D. 161n11

folk theories 46, 49, 79, 145–6, 218–22, 224, 227, 229, 231, 233, 235, 237, 241, 259, 275, 285, 290

Frederick, S. 96n17

Frege, G. 7, 11–12, 16–17, 21, 30, 45, 63, 75, 114, 281, 286

Frege-analyticity 63–70, 72, 116

fuzzy logic 35–7, 39–40, 43–4

Galilei, G. 179, 279

Gallie, W. B. 126n40

Gauker, C. 93n13

Gaut, B. 149n6

Geach, P. T. 21

Gendler, T. 187–8

Gentzen, G. 76, 102, 162

geometry 217

Gettier, E. 9n2, 179–95, 199–200, 203–6, 211, 216–17, 226, 233, 235n16, 239–40, 242–4, 271, 280n1, 306–8

Gettier cases, real life 192–5, 216, 244

Gödel, K. 9n2, 47

Goldberg, S. 98n18

Goldman, A. 148n5, 150n7, 191n10, 236n17

Goodman, N. 143, 150, 244

Grandy, R. 261, 273

Grice, H. P. 50n2, 85, 96, 186n5, 287

Gupta, A. 280

habituals 138–9

Häggqvist, S. 186n5

Halbach, V. 280–1

Hampton, J. 96n17

Harman, G. 100n20, 118n30, 242n19

Harris, P. 103n22, 141

Hawthorne, J. 93n14, 234n14, 268

Hegel, G. W. F. 280

Heil, J. 100n20

Hempel, C. G. 239n18

Heraclitus 278, 289

hermeneutics 14

Heyting, A. 282

Higginbotham, J. 282

Hill, C. 159

Hintikka, J. 21, 212n2

historiography of twentieth-century philosophy 11, 19, 22

history 42, 208–10, 213, 246, 285

history, philosophy of 6

holism

- about content 258–60, 267, 271
- epistemological 91, 97, 125, 131

Horgan, T. 87n9, 218, 220

Horwich, P. 67n19, 79n2

Hughes, G. 302

humanity, principle of 261–2, 273

Humberstone, I. L. 65, 295

Hume, D. 11, 48

idealism 14–17, 49, 212

identity *see* metaphysics of identity; necessity of identity

imagination 107, 136, 143, 147–53, 155, 163, 165–7, 170–1, 174, 177, 179–80, 185, 187–8, 193, 216, 244, 275, 285, 289

inconceivability *see* conceivability

induction 102, 145, 152, 166, 168–9, 190, 225, 252

inference rules 74–6, 79, 83, 92, 104–7, 109, 120n32, 146

inference to the best explanation 5, 137, 146, 166, 169, 182, 208, 224, 236

inferentialism 76, 97, 100n20

intensional isomorphism 127–8

intentionality 14, 18, 260–1, 264, 268–9

internalism, epistemological 5, 209, 270–1

interpretation *see* radical interpretation

intertranslatability 251, 260

intuitionistic logic 32–3

intuitionistic semantics 282–3

intuitions 2–3, 5, 9, 130n43, 168, 178, 188, 195, 199–200, 212n2, 214–21, 223, 228–9, 235–7, 239, 241, 244–5

irrationality *see* rationality

Jackson, F. 121–2, 141

Joachim, H. H. 11n1

Johnson-Laird, P. 106  
 Johnston, M. 17n5  
 Jönsson, M. 96n17  
 judgment skepticism *see* skepticism  
     about judgment  
 justice 190  
 justification 52–3, 63, 70, 77, 81,  
     83–4, 130, 132, 179–86,  
     188–90, 192–5, 197, 205–6,  
     211, 225, 242–4, 270–1  
*see also* understanding–  
     justification links

K (modal logic) 298–300  
 Kahneman, D. 96n, 140n1  
 Kant, I. 19, 180n1, 289  
 Kaplan, D. 21, 65, 69n20, 124n37,  
     127–8, 144, 296n1  
 Keefe, R. 23n1, 28n4  
 Kleene, S. C. 87–9, 95  
 Klein, D. 148n5  
 Klein, P. 261n8  
 Kment, B. 175n15  
 knowledge 5, 7, 20, 31, 39–40, 44,  
     46–7, 49, 52–4, 60–3, 66–7,  
     70–1, 77–8, 81, 84, 110–12,  
     116–17, 119, 124, 129–30,  
     132, 134, 140–3, 145–6,  
     152n8, 164, 167–9, 178–86,  
     188–90, 192–5, 205–7,  
     211–13, 216, 221–2, 225, 239,  
     242–3, 247, 250, 263–77,  
     280  
     and intelligent life 269–70  
*see also* tacit knowledge;  
     understanding–knowledge  
     links  
 knowledge maximization 265–70,  
     272–3, 275–7  
 Kornblith, H. 6n1, 191n10,  
     222n5

Kratzer, A. 142n2, 177n16  
 Kripke, S. 19, 21, 51, 62, 65–9,  
     123, 125, 134, 161, 168, 205,  
     264, 280

Ladusaw, W. 108n27  
 Lakatos, I. 284  
 Lange, M. 175n15  
 Langford, C. H. 69n22  
 language  
     philosophy of 4, 12–13, 18, 46,  
     49, 118, 212, 260, 281–5  
     public 12–13, 89–91, 98–9, 118,  
     122–7  
     of thought 13, 49  
 law courts 41–2  
 laws, natural 135, 141, 146,  
     149

Leibniz, G. 19  
 Leslie, A. 148n5  
 Lewis, D. 8–9, 19, 21, 67, 93n14,  
     100n20, 121–2, 141–2, 144,  
     150–2, 154n9, 156, 159,  
     172–3, 175, 215–17, 220,  
     244n20, 256, 260, 268, 293–5,  
     300, 302–4

liar paradox *see* semantic  
     paradoxes

linguistic competence 2–3, 31,  
     39–41, 48, 52–3, 63, 66–70,  
     73–4, 76–85, 88–92, 94–9,  
     105, 107–12, 114, 116–21,  
     126, 128–33, 147, 158, 168,  
     188, 259, 282–3

linguistic philosophy 2–4, 10–11,  
     51

linguistic turn 2, 10–13, 21–22, 31,  
     49, 53

linguistics 3–4, 10, 50, 110, 112,  
     212n2, 281–2

Locke, J. 53n5

logic 7, 20, 24, 40–1, 45–7, 70–1, 76, 89, 91, 93, 97, 115, 130, 132, 162–3, 281, 284–5, 289–90  
*see also* counterfactual logic; deviant logic; epistemology of logic; intuitionistic logic; modal logic

logical consequence 46, 58, 65, 71, 162, 266

logical constants 76, 105, 107

logical positivism 51, 54, 279

logical truth 25, 31, 63–5, 68, 70–2, 88, 112, 116, 144, 162–3

lottery paradox 95

Lowe, E. J. 93n12

Lycan, W. 93n12

Manktelow, K. 103

Marconi, D. 98n19

Marcus, R. B. 21

Marion, M. 270n11

Martin, C. B. 100n20

Martin-Löf, P. 76

Mates, B. 117

mathematics 4, 6–7, 20–1, 45, 47, 61, 63, 134–6, 168–9, 173, 203, 208–10, 216–17, 236, 246, 264, 274, 279, 282–3, 285–6, 289, 291  
 philosophy of 6

McCulloch, G. 261nn7, 8

McDowell, J. 15–17, 270n11

McGee, V. 28n4, 92–4, 99, 280

McGinn, C. 261n7

McKinsey, M. 169n

McLaughlin, B. 28n4

meaning *see* semantics; skepticism about meaning; synonymy

mental models 106–7

mereology 15

metaconceptual questions 3, 15n3, 21, 23, 29–31, 36, 41–6, 48–9, 76–7, 211, 284

metalinguistic questions 21, 23, 26–8, 31, 36, 41–6, 48–9, 59, 86, 90, 109–10, 117, 284–5

metalogical questions 40–1, 43

metaphysics 18–20, 46, 48–9, 51, 87n9, 164, 211, 218–19, 221–3, 236, 259–62, 269n11, 281, 284–5  
 of analytic truths 52–4, 63, 71–3  
 of identity 123–5, 211, 214, 280  
 of words 27–8, 127  
*see also* modality, metaphysical; time, philosophy of

metasemantic questions 71–2

methods, philosophical 3, 5, 7, 10, 12, 14, 18, 46, 164, 179, 242, 244, 278, 285, 287

mind, philosophy of 13, 18, 49, 212

mind-independence 134–5

modal logic 62, 135, 156–61, 184, 187, 201–2, 280, 296, 298–304

modal-analyticity 60–3, 65

modality, metaphysical 4, 19, 25, 51, 63, 65, 67, 134–6, 155–65, 167, 170–1, 174, 178, 187, 204–7, 219, 227–9, 249, 256, 273, 280, 300–1  
*see also* epistemology of metaphysical modality; modal-analyticity; quasi-Frege-analyticity; skepticism about metaphysical modality

modality, physical 135, 161n11

modus ponens 44, 92–5, 103, 139, 152n8, 160, 225, 254–5, 300–1

molecularism about content 258–9  
 Montague, R. 281  
 Moor, J. 93n12  
 Moore, G. E. 12, 21, 218  
 moral philosophy 49, 164, 185, 206  
 Müller-Lyer illusion 216  
 Mumford, S. 100n20

Nagel, T. 260  
 naïve comprehension principle 216–17  
 natural deduction 76, 86, 95, 102, 106, 162  
 natural kind terms 20, 61, 76, 78–82, 109–10, 123–4, 129, 164, 168, 170, 206, 264  
 natural sciences 1–4, 19, 42–3, 46, 135–6, 179, 191, 194, 208–10, 212–13, 220–3, 239, 241, 244, 246, 275–6, 278–80, 287  
 naturalism 2, 18, 49–50, 220, 251, 275  
 naturalness 122n34, 256–8, 266, 268  
 Neale, S. 199n17  
 necessity *see* modality, metaphysical; modality, physical  
 necessity of identity 161, 171, 174, 280  
 NECESSITY (principle) 156, 159–60, 170, 172, 202n20, 297–8  
 negative polarity 107–9  
 Newstead, S. 104  
 Nichols, S. 148, 188n8, 190n9  
 Nolan, D. 93n13, 171  
 Norton, J. 187n7  
 Nozick, R. 118n30, 152n8, 194–5, 234n14, 249n2  
 nominalism 19  
 Oaksford, M. 103n21  
 O'Brien, D. 105n24  
 obviousness 272–3  
 Olson, J. 140n1  
 Osherson, D. 6n1, 96n17  
 Over, D. 93n12, 103, 105n24, 142  
 paleo-pragmatism 278  
 Partee, B. H. 282  
 Peacocke, C. 13, 74, 76, 81n5, 95n16, 98n18, 149n6, 177n16  
 pejorative terms 80n4  
 perception 1, 5, 18, 47, 103–4, 107, 120n33, 133, 136, 143, 147–50, 152, 165–70, 186n5, 189–92, 216–19, 220–1, 223, 225–7, 235, 238, 244, 248–50, 263, 266–7, 273, 275–6  
 perceptual demonstratives 263–4, 266–7  
 phenomenology 14, 217  
 physicalism 67  
 physics 18, 21, 45n14, 179, 221, 231, 274–5, 285  
 physics, philosophy of 6, 18, 135  
*see also* philosophy of time  
 Plato 180  
 political philosophy 49  
*see also* justice  
 possibility *see* conceptual possibility; epistemic possibility; modality, metaphysical; modality, physical  
 POSSIBILITY (principle) 156–7, 159–60, 172n14, 187, 202, 205, 297–8  
 possibility, easy 177–8  
 possible worlds 67, 141–2, 151, 156, 159n10, 172, 176, 186, 254, 257, 280, 305–7  
 postmodernism 14

Prawitz, D. 76  
 preface paradox 95  
 presupposition 86, 88  
 Price, H. H. 11n1  
 Prichard, H. A. 270n11  
 Priest, G. 94, 126  
 Prior, A. N. 79  
 privileged access to mental states 169n13, 236–7, 245  
 progress, philosophical 7–8, 66, 279–80, 284, 286–7, 291–2  
 proof 7, 9n2, 32, 45, 134, 173–4, 203, 208–10, 283  
 propositions, Russellian 15, 66–7, 75  
 psychologization *see* evidence, psychologization of  
 psychology 7, 212, 216, 236, 247, 285  
 Pust, J. 236n17  
 Putnam, H. 32, 69, 91, 94, 97–8, 109, 123, 264

quantification  
 into modal contexts 161  
 into sentence position 159, 296  
 quasi-Frege-analyticity 70  
 question-begging 214  
 Quine, W. V. O. 19, 21, 26n1, 39, 50–2, 59n9, 63–4, 85, 91, 97, 127, 272–3

radical interpretation 260–1, 272–3  
 Ramsey, F. 121–2, 150n7  
 rational reflection 169n13  
 rationalism 1–2, 4, 130n43, 216  
 rationality 91–2, 114, 126, 203, 239, 246, 252, 256, 262, 265, 270  
 Rawls, J. 21, 244  
 realism / anti-realism debate 281–4, 287–8, 290

reasoning, psychology of 100, 102–9, 112, 145  
 reference 77, 110, 122n34, 124–5, 258–9, 261, 263–71, 273, 281  
 causal theories of 258–9, 263–4, 271  
*see also* direct reference  
 reference failure 20, 78–82, 129, 184  
 reflective equilibrium 5, 244–6  
 register 129  
 relevance logic 94–5  
 reliability 3, 6–7, 62, 83–4, 103, 105, 137, 146, 149–50, 155, 164–6, 171, 194, 213, 224, 238–9, 245, 268, 281, 284  
 rigor 7–8, 46, 65, 215, 286, 288–9  
 Roberts, C. 308  
 Roese, N. 140n1  
 Rorty, R. 10, 12  
 Rosen, G. 136  
 Russell, B. A. W. 12–13, 15, 21, 45, 75, 180n2, 217  
*see also* propositions, Russellian  
 Russell's paradox 217

S4 (modal logic) 301–4, 307  
 S5 (modal logic) 62, 135, 160, 201–2, 294, 301–4, 307  
 Sainsbury, R. M. 177n16  
 Salmon, N. 66n, 303  
 Schaeken, W. 103n21  
 Schechter, J. 162  
 Schroeter, F. 123n36  
 Schroeter, L. 123n36  
 Schroyens, W. 103n21  
 science, philosophy of 6, 241–2, 244  
 semantic paradoxes 79, 83, 94, 280–1

semantics 4, 6, 20, 31, 37–40, 45–6, 50, 57–8, 63–4, 70–2, 87–90, 93, 97, 99, 102, 106, 110, 115, 119, 127–9, 131–3, 141–2, 150, 156, 158, 162, 169n13, 173, 175–8, 196, 199n17, 266, 281–5, 294–5, 301–8

*see also* assertability-conditional semantics; externalism, semantic; intuitionistic semantics

sense, Fregean 16–17, 30, 75

set theory 169, 213, 216–17

Shapiro, D. 103

Shope, R. 168, 180n1, 188n8

Sides, A. 96n17

simulation 147–53, 155, 166, 169–70, 188, 216

and expectation-forming capacities 148–51, 153

Sinnott-Armstrong, W. 93n12, 214n

skepticism 3, 81, 146, 155, 182, 193, 211, 218–27, 230–2, 234–6, 238–41, 248–50, 252, 261, 265, 268, 271, 275, 277

about counterfactual conditionals 141, 155, 165

about judgment 220–31, 232–5, 237, 239–41, 250–2, 273–6

about meaning 52, 60, 71, 127

about metaphysical modality 60, 136–7, 162, 164–5

about philosophical thought experiments 179, 188, 191, 194–5

about reason 238–40

probabilistic treatment of 227–34

Smart, J. J. C. 21, 48n1

Soames, S. 87, 111n28

Sober, E. 50n2

social dimension of philosophy 1, 7–8, 68, 90, 112, 114, 280, 286–7, 290–1

Sorensen, R. 120n, 186n5

sorites paradox 33, 120n33, 122

Sosa, E. 190n9, 216

Stalnaker, R. 21, 67, 72n24, 88, 138, 141, 151–2, 154n9, 156, 159, 175, 300, 304

Stanley, J. 93n14

Stanovich, K. 104

stereotypes 109

Stich, S. 148, 188n8, 190n9, 244n20

Stine, G. 93n14, 234n14

Strawson, P. F. 20–1, 50–1, 85

subjunctive conditionals *see* counterfactual conditionals

supervaluationism 28n4, 43, 96, 110, 122n34

supervenience 194, 260–1

Sutton, J. 182

synonymy 13, 29, 50–2, 63, 66–9, 90, 116–18, 127–9, 160, 183

syntax 101, 104, 107–8, 285

system 1–system 2 distinction 104–5, 107

T (modal logic) 300–1

tacit knowledge 110–12

Tappenden, J. 53n4

Tappolet, C. 58n8

Tarski, A. 46, 64–5, 79, 282

testimony 40n11, 118n31, 124–5, 152, 167, 170, 192, 220, 225, 250, 275

Thales 278

theoretical terms *see* natural kind terms

thought experiments 19, 153, 164, 179–81, 185–95, 200–7, 244, 246, 279, 308

three-valued logic 34–9, 43–4, 87–9, 95

time, philosophy of 6, 19, 49, 87n9, 211, 280, 290

total evidence, requirement of 239, 274, 276

training

- legal 191
- mathematical 286, 289
- philosophical 7–8, 191, 286, 290
- scientific 191

truth

- formal theories of 280–1
- in virtue of meaning 52, 58–9, 61
- see also* disquotation; semantic paradoxes; understanding–truth links

truth-indicativeness 227–31

truthmakers 59–60

Tversky, A. 96n17, 140n1

two-dimensionalism 169n13

understanding *see* conceptual competence; linguistic competence

understanding–assent links 74–86, 92, 99, 110–13, 116, 120–1

understanding–disposition-to-assent links 100–2, 113, 116, 120–1

understanding–justification links 81, 83–5

understanding–knowledge links 77–8, 80, 82–5

understanding–truth links 78–81, 83

unity questions 123–5

vagueness 23–4, 28, 31–2, 35–41, 43, 50, 52, 87–9, 94–6, 98, 105, 110, 120n33, 122n34, 124n37, 175–6, 218, 262, 281, 288–90

Vahid, H. 242n19

validity *see* logical consequence

van Inwagen, P. 19, 21, 215–18, 220

van Rooij, R. 108n27, 199n17

verificationism 258–60, 278, 282

Viale, R. 6n1, 96n17

Vico, G. 149n6

von Fintel, K. 176

Wason, P. 103

Weatherson, B. 235n16, 268

weak centering axiom 300–1

Weber, K. E. M. 149n6

Weinberg, J. 188n8, 190n9

West, R. 104

Wiggins, D. 19–21

Williams, B. 21, 149n8

Wilson, J. C. 11n1, 269–70

Wittgenstein, L. 8, 12, 19, 21, 37, 53–4, 65, 125

words *see* metaphysics of words

Wright, C. 136, 282

Yablo, S. 167n12

Zimmerman, D. 19n6